
MEASUREMENT FOR GUIDANCE

MEASUREMENT FOR GUIDANCE

Exploration Series in Education
Under the Advisory Editorship of
JOHN GUY FOWLKES

Measurement for Guidance

JOHN W. M. ROTHNEY

University of Wisconsin

PAUL J. DANIELSON

University of Arizona

ROBERT A. HEIMANN

Arizona State College

HARPER & BROTHERS



PUBLISHERS, NEW YORK

If I want to understand an individual human being, I must lay aside all scientific knowledge of the average man and discard all theories in order to adopt a completely new and unprejudiced attitude. I can only approach the task of *understanding* with a free and open mind, whereas *knowledge* of man, or insight into human character, presupposes all sorts of knowledge about mankind in general.

CARL JUNG

CONTENTS

EDITOR'S INTRODUCTION	ix
PREFACE	xi
I. TESTING IN GENERAL AND IN THE HIGH SCHOOL COUNSELING PROGRAM	1
The Counseling Task—The Counseling Program—The Magnitude and Scope of Psychological Testing—The Development of Tests as Aids to Counseling—Testing in Manpower Utilization and in Personal Development—The Place of Tests in Counseling—Summary	
II. VARIETIES AND SOURCES OF TESTS	27
Types of Tests—Sources of Test Materials—Checking on the Test Publisher—Summary	
III. CRITERIA OF TEST SELECTION	49
Validity—Validity for Counseling—Reliability—Norms—Administration of Tests—Scoring of Tests—Mechanical Considerations—Summary	
IV. TEST SCORES: ETIOLOGY AND INTERPRETATION	116
Selection of a Test—Factors Influencing Test Scores—Summary	
V. THE USE OF STANDARDS IN TEST SELECTION	151
Technical Recommendations for Psychological Tests and Diagnostic Techniques—Application of Recommendations to the Cooperative School and College Ability Tests—Summary	
VI. RECORDING AND REPORTING TEST SCORES	204
Methods of Recording—Summary	
VII. COMBINING TEST SCORES WITH OTHER DATA	226
Combining Test and Clinical Data for Counseling—Information Provided by Personal Data—Contradictory Evidence:	

Cases—"That Depends"—Individualized Test Interpretation
—Summary

VIII. PERSONALITY QUESTIONNAIRES AND INTEREST INVENTORIES 282

Limitations of Short-Cut Methods—Attempts at Justification of
Short-Cut Methods—Objections to Short-Cut Methods—Valid-
ity of Interest Inventories—Reasons for Lack of Validity—
Validity of Personality Appraisal Techniques—Validity of
Projective Methods—Reliability of Interest and Personality
Measurements—Norms of Personality and Interest Inventories
—Suggestions for the Counselor—Summary

IX. THE FUTURE 320

Some Possible Developments—Some Basic Problems in Test
Development—Current Signs of Progress—Summary

APPENDIX: ACTIVITIES REPORTS 353

INDEX OF NAMES 371

INDEX OF SUBJECTS 375

EDITOR'S INTRODUCTION

Schools, students, and scholarship are in the limelight of the press, the pulpit, and home as never before in the United States of America. The importance of what happens in educational institutions *from the first day of kindergarten through the last day of formal education* is being emphasized in all quarters. Statesmen as well as schoolteachers reflect an inspired determination to make sure that adequate funds are provided to offer the quality as well as the kind and amount of educational opportunity needed for the optimum continued development of our country.

Furthermore, it is recognized that learning must continue throughout all adulthood as well as in childhood and early adulthood. Increased offerings for "adult education" are being made by both public and private institutions.

The circumstances just cited are indeed gratifying, comforting, and inspiring, particularly to all those engaged in educational activities. On the other hand, these circumstances are also sobering and challenging on many counts, but especially in connection with the professional counseling of the individual boy or girl in the elementary or secondary school, the college and university student and, indeed, parents and grandparents.

The need for and confidence in professional counseling is evidenced in many ways. Specially trained workers in guidance and counseling are found in all types of educational institutions. Many commercial agencies also engage in advising men and women as to what they "should do" particularly with reference to the "job" they long to get.

Intelligent and valid counseling is possible only if the counselor is as familiar as possible with the abilities, habits, strengths, and weaknesses of those who are counseled—in short, the counselor must know the counselee. One of the bases most commonly used

by counselors in learning about their counselees is the standardized test.

It is therefore obligatory that professional counselors be wise in the selection of tests, skilled in their administration, and show deep and discriminating insight in the interpretation of test results. In short, a thorough and discriminating familiarity with measurement is essential for effective counseling.

This refreshing and stimulating book deals with the functions, appropriateness, selection, use, recording, and interpretation of tests with respect to their value in the counseling of individuals. Weaknesses as well as strengths of tests are treated with the delicate discrimination and judgment that come from years of work with tests in relation to individual counseling. To the writer the highlight of this volume is the last chapter, entitled "The Future."

This concluding chapter presents a moving challenge and also points the way towards more valid tests and improved use of them.

All those who work with tests, but especially counselors, will experience pleasure as well as professional improvement from using and pondering over this work.

JOHN GUY FOWLKES

PREFACE

This volume is intended for use in the preservice and in service training of counselors and others who attempt to use tests in the counseling of *individuals*. The emphasis is on the practical application of measurement in counseling for the large group of guidance workers who are familiar with elementary concepts of measurement but who are not yet ready for advanced technical work in theory. It is assumed that such persons will have had a first course in tests and measurements and will have developed some competence in the interpretation of elementary statistics.

Most tests are designed to be used with groups but many authors of tests claim that their instruments may be used successfully in the educational and vocational guidance of individuals. Thus counselors, if they are to employ tests in their work, are required to use instruments that may serve their stated primary purposes well but may fail completely to meet their second objective. The values and limitations of such instruments in counseling of individuals are considered fully in this volume.

Discussion of several widely used tests of individuals has been purposely eliminated from this book. The Stanford-Binet and Wechsler-Bellevue tests, for example, are not considered here because they have been thoroughly discussed by their authors in volumes devoted entirely to their construction and use. In addition there seems to be sufficient evidence to indicate that they do not usually contribute enough per unit of cost in the counseling of most of the subjects a counselor meets in schools to justify their use, although they may do so in the clinical study of unusual individuals. Similar reasons may be cited for the omission of discussion of those individual tests that purport to measure such factors as finger dexterity and mechanical manipulation. Counselors who plan to use such instruments must take special training in their

use that goes beyond the scope of this volume, although some of the basic problems in the interpretation of scores obtained from them are similar to those that are discussed in the following pages.

Samples of items extracted from a test and reproduced in a book cannot indicate its nature and value. Study of a whole test and the manual accompanying it is essential if the counselor or counselor-in-training is to determine whether the test may be useful in his work. It is essential, therefore, that a file of tests and measurements be made available for use by students while they use this volume. For intensive study of some tests it seems desirable to have tests and manuals available for each member of a class. Samples are available without cost or at reduced rates for instructional purposes from most publishers of tests.

In many places throughout this volume the authors have been extremely critical of some common *practices* in measurement while they have accepted many of its basic principles with enthusiasm. The types of tests and practices that are criticized here are only samples of those commonly used. Many millions of tests are used annually by counselors despite the fact that their authors present no evidence or inadequate evidence that they can be used effectively in the counseling of *individuals*. The waste that may result and the wrongs that may be done to counselees are matters for serious consideration by those who work in the field of guidance. It is hoped that the criticism and suggestions presented in this volume may result in a demand by counselors that better tests be produced. It is believed that the use of such improved instruments will increase the effectiveness of guidance workers in their service to youth who meet real personal problems and who must make decisions that are crucial to themselves and to society.

JOHN W. M. ROTHNEY
PAUL J. DANIELSON
ROBERT A. HEIMANN

August, 1958

MEASUREMENT FOR GUIDANCE

CHAPTER I

Testing in General and in the High School Counseling Program

Throughout this volume emphasis will be placed on the use of tests in the school counseling programs. The authors believe that the employment of tests by counselors differs from the use made by others in terms of concepts, technical requirements, and applications. Their use for the special purposes of counseling is the major concern of this volume. In the opening chapter a brief overview of the general scope of testing will be presented and some of the circumstances that make testing in a counseling situation, as defined by the authors, different from those found in other applications will be discussed.

THE COUNSELING TASK

"I don't know just what I want to follow as an occupation; can you give me a 'vocational test' that will give me the answer?"

"I thought I would like to take some tests to see if I can qualify for my current choice of college curriculum."

"I wish you would give my son a vocational aptitude test because

he can't make up his mind about what he wants to do after his graduation."

Every school counselor has heard these or similar questions many times. And many times, perhaps, he has resisted the temptation to answer abruptly, ". . . No, I can't and only a charlatan would. . . ." Yet, if he did so, he might be too harsh on the person making this common request, for the applicant usually believes that he is taking the right step in requesting some "scientific" measurement of his aptitudes and abilities. This procedure seems very much the thing to do in the middle of the twentieth century, when we seem to be living in the age of the specialist, when we have rationalized many of our procedures and behaviors, and when we seek "expert" advice about our everyday affairs.

The counselor, a person with specialized training and a rich and varied background, is concerned with all aspects of the performance and behavior of his counselees. As he assists them to develop their attitudes and feelings in relation to vocational, educational, or personal choices, he tries to keep before them the multiplicity of choices that are theirs to make and the probable resulting consequences. He is deeply concerned about them and he communicates this concern as he works on the development of a counselee's understanding of himself and the making of decisions.

And these decisions that a mid-twentieth-century youth must make have dimensions never before so complex or pressing in western civilization. While still in high school he must choose with discernment from among thousands of occupations, more than one hundred different college training curriculums in some 2,000 colleges, and from some three dozen ways of meeting his national service obligations. These increasingly complex choices are his to make. At one time these choices could be made with the help of such highly integrative social institutions as the family, the stable small town unit with rather fixed, crystallized class and occupational structure, and the traditional academic secondary

school. These institutions are rapidly taking new and strange shapes and forms.

The youth may feel vague unrest and some anxiety over the spectre of automation in industry with its increasing threat of technological displacement and possible unemployment. He may see increasing concentration of industrial power by big industry with resultant and discomfoting impersonalization. He may see society increase its demands for the highly trained specialist who must concentrate his technically acquired talents in narrower and narrower fields.

The difficulty of making decisions under such circumstances may be compounded by an overspreading aura of anxiety, for this mid-twentieth-century youth finds his very society torn between conflicting world-wide ideologies and his world divided into two armed camps. The impact of two world wars, a major world-wide depression, and the threat of the H bomb have left their stamp upon his parents and his teachers and they may communicate some of their uneasiness to him.

It is in terms of this broad social base that the youth approaches his counselor in a school setting for help in choosing among the best bets or the most likely avenues of decision in relation to his future. It is in this framework that the counselor tries to bring a little concern for individuality to this same bewildered youth. And although the educational system attended by the young mid-twentieth-century youth has given lip service to the idea of individual differences for several generations, he is likely to find it *in reality* only in the intimacy of a counseling relationship. This deep personal concern for each young person as an individual is the prime requisite of the recently developed school counseling service and the major distinguishing characteristic of the counselor. It comes as a healthy antidote to the spreading depersonalization of American secondary education and its bell-curve-dominated, fractionized, compartmentalized schools with their diverse curricula. The counselor stands as the one person—central and warm—to whom

the youth can relate. The counselor may become a stable integrative force in the seas of uncorrelated, unrelated experiences that the youth often finds in the high school of today. But it is when he attempts to do something to show his deep personal concern for each individual by helping him to help himself that the counselor finds his testing instruments particularly inadequate and ineffective.

THE COUNSELING PROGRAM

It is not the purpose of this volume to describe counseling programs in general but some elaboration of common underlying concepts and frequently stated assumptions of such programs are presented below to indicate the settings in which tests are used by counselors.

The major objectives of guidance programs at the secondary level are usually stated in terms similar to these:

1. "The provision of assistance to *all students* in order that they may recognize their limitations and potentialities, to the fullest possible extent, and to utilize this knowledge and development in planning their school and post-school careers.

2. "To coordinate the efforts of home, school, and community to assist students toward the goals named above."¹

The nature of the individual student is such that the successful attainment of these objectives can be realized only to the extent that counseling becomes an individualized affair. This point is made clear by Rothney and Roens in their statement that:

Counseling must always be an individualized affair. . . . The word "always" is used advisedly for the foundation of counseling is found in the fact that there are personal choices to be made. In many cases there are similar situations and patterns of development which require similar choices, but, in the last analysis, there must be some one person who accepts the responsibility of helping this particular indi-

¹ John W. M. Rothney and Bert-A. Roens. *Guidance of American Youth*. Cambridge: Harvard University Press, 1950, p. 191.

vidual to analyze his unique personal problems. To such situations someone must bring particularized knowledge obtained from records, observations, and tests, and someone must interpret it. Someone must answer the student's specific questions, and someone must raise particular questions that he may not have raised about himself. Someone must interpret to each student separately the specialized educational and vocational implications which he, because of his lack of experience and knowledge, is unable to recognize, and someone must help each student to appreciate the social and domestic circumstances of his particular characteristics and situation. Someone with quick personal perceptions and a sympathetic interest in human difficulties must help a student to help himself when he finds that he is confronted with problems beyond his power to solve. . . . It is these personalized tasks, then, that the counselor, who has only a token teaching assignment and who has had specific training, will undertake.²

In carrying out this assignment it is assumed that administrative arrangements will provide the time and facilities to enable the *counselor to confer with his subjects frequently over the years that they are in school.* It is also assumed that enough rapport has been established so that the students feel that assistance will be provided if it is sought.

THE ROLE OF THE COUNSELOR³

The central figure, the person, the "someone" referred to in the section immediately above is the counselor who, in assuming his

² John W. M. Rothney and Bert A. Roens. *Counseling the Individual Student*. New York: Dryden Press, 1949, p. 4.

³ *In the present stage of development of counseling, any attempt to define the task of the counselor will have to be quite arbitrary. The questions of who is the counselor and what he does have been disturbing for many years, and most attempts to spell out the task of the counselor have resulted either in something quite vague or in a synthesis so general as not to be especially helpful. Those who have been engaged in something called "counseling" will recognize the problem; those in training will probably experience it in future moments of introspection if, indeed, they have not already done so in their search of the literature.*

At best, it would appear that an answer to the question "What is the task of the counselor?" must be prefaced with ". . . that depends." It will depend on the ex-

role, will work in a face-to-face and one-to-one relationship with the counslee in which he will "consciously attempt by verbal means to assist [the student] in modifying attitudes and other behavior with respect to educational, emotional, and vocational issues."⁴

QUESTIONS TO BE CONSIDERED

During interviews the counselor will be required to perform many different functions. Some of the verbs that may be used to describe his activities are interpret, inform, listen, describe, compliment, encourage, refer, demonstrate, provide, assist, and confer. Any list of this kind cannot be complete because special circumstances will require special action. He will do these things as he seeks to fulfill the objectives of his program. As the counselor goes about his duties he will find that the problems raised and the issues discussed will tend to run the gamut of human difficulties as his subjects try to find their way among the forests of their own desires and the roadblocks that society puts up against them. He will find that he must answer the questions of students, parents, school personnel, potential employers, and personnel of institutions for advanced training. He will find that he must help these individuals to find answers to their own questions and he will realize that he must raise questions that these persons have not known enough to ask. He will discover, too, that he must seek answers to questions that arise to him as he works with these persons.

tent of training of the counselor, the "school of thought" under which that training was obtained, the "philosophy" of the administration under which the counselor is working, the facilities at his disposal, the referral and allied agencies with which he can work, whether the counselor is THE guidance program or a "service plus" among many other services that may be provided. Because of the many variables involved it is not difficult for the reader to see why no wholly inclusive and mutually acceptable definition of duties has been evolved.

⁴ John W. M. Rothney and Paul J. Danielson, "Counseling," *Review of Educational Research*, April, 1951, 21:132-139.

The questions that are asked, or sometimes just implied, are so numerous, unique, and varied that they defy listing or classification. The questions given below are some that counselors often meet, but they must be considered only as samples.

QUESTIONS BY STUDENTS

1. Am I capable of undertaking certain training successfully?
2. Why can't I do this work well? Why am I having trouble with this course?
3. Do I have any particular strengths or weaknesses?
4. Could I get into officer training when I get in the armed services?
5. What are my chances of winning a scholarship?
6. Do you think I could pass the draft deferment test if I go to college?
7. My teachers say I am not working up to my ability. What do they mean?
8. Why can't I learn things in some courses when I can in others?

QUESTIONS BY PARENTS

1. Is my child capable of undertaking a certain kind of training successfully?
2. Why is my child having difficulty in his training?
3. Could my son be successful as a mechanic (or any other specific occupation)?
4. In view of my child's health would it be wise to plan for training beyond high school?
5. Should my child be accelerated? Should he repeat some of his work?
6. Is my child better fitted for one kind of work rather than another?

QUESTIONS BY SCHOOL PERSONNEL

1. Why is this student not working up to capacity?
2. Why is this pupil having difficulty in my course?
3. Wouldn't it be better to guide this student out of a particular program of study?

Measurement for Guidance

4. Is this pupil actually unable to do this work or is it just a case of not trying hard enough?
5. Shouldn't this student be in a fast section? a slow section?
6. Is this pupil really as dull as he seems?

**QUESTIONS BY PERSONNEL OF
INSTITUTIONS FOR ADVANCED TRAINING**

1. Does this applicant show promise of success in our institution?
2. Has this student shown any particular strengths or weaknesses?
3. Is this pupil well enough prepared to undertake a particular course of study?
4. Has this student developed good habits of study?
5. Is this student worthy of scholarship aid?
6. Is the applicant sufficiently mature to undertake this training?

QUESTIONS BY EMPLOYERS

1. Would this applicant do better in one phase of our work (mechanical) than in another (clerical)?
2. Can you provide us with information that will assist us in the placement of this individual within our organization?
3. Our training program within our company provides the following opportunities. (They are described.) In which of these do you think we ought to place him?
4. How does he get along with others?
5. We are interested in potential leadership—How will he respond to training?

QUESTIONS BY THE COUNSELOR

1. In view of this student's stated choices, what educational program would seem best for him?
2. During the time left in school what are the probabilities that his performances may change?
3. What explanation can be found for the apparent inconsistencies in this student's record?
4. Is this student's visual condition such that he should not undertake a lengthy program of training that requires much reading?

SOURCES OF ANSWERS TO THE QUESTIONS

As the counselor sets out to find the answers to such questions, and to help others to obtain them, their variety and complexity demand that he turn to many sources, procedures, and techniques. He will begin his search with acceptance of the fact that there is no single procedure that will guarantee success, no one source that will be infallible, and no particular technique that can always be applied. Except in the simplest situations the counselor will seldom find that there is a clear-cut course of action. He will usually find that he and his questioner must embark jointly on an enterprise that will be complex and time-consuming over a considerable period.

It is impossible in the current status of our knowledge about human beings to decide, before conferring with a student, what information about him needs to be obtained so that his questions can be answered. Rothney and Roens⁸ have set up some *general* guides for the collection of information about students. They include emphasis on the need for finding out what is particularly important to each counselee, the need for longitudinal data about him, the necessity of having information about his cultural milieu, the attitude of holding final conceptualization of the person in abeyance till all possible sources of information have been examined, and the willingness to appraise all sources of information thoroughly before they are used. These, however, are general guides and the specifics must depend on circumstances and conditions. To answer the questions of a counselee who comes initially because of vocational indecision the counselor may have to consider, among other things, his subject's health, previous performances and experiences in related areas, the financial and social circumstances of his home, the opportunities for training and employment in the areas considered, his usual behavior and signifi-

⁸ Rothney and Roens, *op. cit.*, pp. 48-64.

cant variations from it, his enthusiasms, his social adjustability, and *anything* about him as a *particular* person that may assist in resolving the indecision.

The sources to which the counselor may turn will then be many and varied. No classification of such sources can ever be adequate but the general listing of them below may serve to point out their scope. In general the following may be used.

1. *Interviews with:*

The counselee

Parents

Peers

Employers

2. *Study of the counselee's records:*

School

Employment

Health

Activities in the community

3. *Examination of past performance of the counselee:*

Personal documents

Productions as in art, music, shop

4. *Files of occupational materials:*

General information

Local data

5. *Source books:*

College catalogues

Handbooks

6. *Test performances*

THE MAGNITUDE AND SCOPE OF PSYCHOLOGICAL TESTING

There has probably been no accurate count of the number of psychological tests administered annually, but it is evident from some of the figures reported that we are rapidly reaching the point at which few persons will have escaped their influence during their

lifetime. The following estimates will give the reader some idea of the magnitude of "operation testing" as it has developed in recent years.

In a recent brochure prepared in the observance of the fiftieth anniversary of a major test publisher it was estimated that "upwards of 75,000,000 standardized tests are given annually in the schools." * Estimates of another test publisher [†] place the figure at more than 100 million, or an average of three tests per pupil at the elementary and secondary school levels. Recently a representative of a third major test publisher displayed with some pride, while visiting one of the authors, a facsimile check in the amount of some \$56,000 which represented the cost of a test order from a single large city school system.

The magnitude of educational testing can be noted further in excerpts from the annual report of the Education Testing Service.[‡] In its report for the year 1954-55, it was indicated that some 171,644 individuals were tested by the College Entrance Examination Board between August 11, 1954, and May 21, 1955; that some 4,000 were tested for admission to the United States Air Force Academy in March, 1955, 14,000 were tested for the General Motors National Scholarship plan, and 22,300 civilian candidates were examined for the Naval Reserve Officer Training program. Examinations were prepared and administered to applicants to the United States Military Academy, the Coast Guard Academy, and the United States Merchant Marine Academy. The Coöperative Test Division of the Educational Testing Service, through its Freshman and Sophomore programs, planned, scored, and interpreted 532,426 college-level tests between the fall of 1953 and spring of 1955. In the year ending June 30, 1955, 51,789 tests were administered as a part of the Graduate Record Institutional

* *Standardized Testing—an Adventure in Educational Publishing*. Yonkers, N.Y.: World Book Co. No date.

† Lyle M. Spencer, *Guidance Newsletter*. Chicago: Science Research Associates, January, 1953.

‡ Educational Testing Service. *Annual Report*. Princeton, N.J., 1954-55.

Testing Program. National Teacher Examinations were administered in 1954-55 to 9,165 candidates, and the Medical College admission test was given to 12,646 candidates. Between July 1, 1951, and June 30, 1955, 1,144,777 tests were given to 354,818 candidates in the total ETS Supervised Testing Program, Special Testing Programs, and Institutional Testing Programs. These figures do not represent the tests sold outright for use by agencies that did their own testing. The figures suggest something of the number of persons involved in testing but, perhaps of greater significance are the implications of the purposes for which they were given.

Business and industrial concerns have become major consumers of tests. Because many industrial tests are custom-designed and not distributed through usual channels, an estimate of the total industrial testing operation is difficult to make. Some of the numbers used may be obtained, however, from such sources as the following. In an article dealing with the increasing use of tests in industry, Whyte⁹ reported that 1,000,000 copies of a personality questionnaire were sold by a single distributor in one recent year. *Fortune*,¹⁰ in a review of testing done by management, reported that one "Aptitude Index" devised by the Life Insurance Agency Management Association had been administered to half a million prospective salesmen by 174 life insurance companies since 1938. The survey further indicated that some 560 industrial tests were available. In 1950 the Psychological Corporation sold more than \$600,000 worth of services to industry. In ten years of operation (up to 1950), the Klein Institute for Aptitude Testing sold 100,000 test batteries to more than 800 companies. The importance of test results to the individuals involved is implied in one example, where the survey reports that "Social Research, Inc. has sold the TAT (Thematic Apperception Test) to about 75 customers, some

⁹ William H. Whyte, Jr. "The Fallacy of Personality Testing." *Fortune*, September, 1954, 50:117.

¹⁰ "The Tests of Management," *Fortune*, July, 1950, 42:92-96.

of whom will not make an executive change without administering the test." ¹¹

Many "consulting firms" have found a ready market for psychological testing in industry. Many industries have employed their own psychologists to aid, through testing, in the selection of employees. They use some of the hundreds of tests available commercially or devise tests to meet their own needs.

The figures given above could be supplemented by those from other major users of psychological tests. The reader may be one of over 18 million men who, during and since World War II, took two or three tests, such as the Army General Classification Test, the Radio Operator's Aptitude Test, the Mechanical Aptitude Test, or the battery administered to enlistees in the navy, in the air corps classification program, and those that may have accompanied applications for Officer Candidate Schools, or other special assignments. Other readers may be among those presently in college rather than in the armed forces because of their performance on the Selective Service Exam. Some readers, while they were seniors in high school, or as registrants for work, may have participated in the testing program sponsored by the United States Employment Service,¹² in which some half a million youth are tested each year. And there are certainly among the readers of this volume persons who were administered a battery of tests when they made application for training under the GI Bill.

Many more samples of the use of tests could be presented here, but enough have been given to indicate the extent to which psychological testing is touching the lives of great numbers of individuals. While the phenomenal growth in the testing movement has not been a recent development, the greater part of the movement has taken place since World War I and particularly since the beginning of World War II. With this overview of the general

¹¹ *Ibid.*, p. 104.

¹² Beatrice J. Dvorak, "The General Aptitude Test Battery." *Personnel and Guidance Journal*, November, 1956, 35:145-156.

growth of the testing movement in mind we may now turn to the more specific study of the growth of the use of tests in counseling.

THE DEVELOPMENT OF TESTS AS AIDS TO COUNSELING

Originally, testing, and more specifically mental testing, were used for classifying individual children as feeble-minded or normal. Within the first 25 years of the twentieth century, the mental testing movement adopted the same mass procedures and large-scale standardization that became common to many other processes on the American scene. Faced with the need to classify 2 million soldiers under the emergency conditions of World War I, group mental tests and group adjustment inventories were developed. The scores derived from the group tests approximated the scores from the individual mental tests but they also lost the sensitivity of the slower, more time-consuming methods.

As an aftermath of the war, group tests of all sorts blossomed forth in almost every conceivable area and the mass testing movement of the 1920's was born. Group tests of mental level and achievement were widely used in education. Industry that had long undergone a specialization and standardization process attempted to find in the mental testing movement a ready answer to its needs for easy and rapid employee selection. The concept of "finding the square peg for the square hole" became popular, and the psychological aptitude test was proclaimed as the best, and certainly the speediest, instrument for appraising the peg. Tests were developed for such occupations, among others, as streetcar conductors, gum-wrapping machine operators, and drill press operators. A job applicant in industry was just as likely to face an Otis Intelligence Test as a high school youth who was to be assigned to a fast or slow section of American history.

Basic to this new concept of management in industry and the measurement movement in education was the actuarial concept that through the testing of large numbers of people and determi-

nation of their average performances, standards could be set up. From these performances cutting scores were derived and a candidate who scored below them was regarded as a poor risk for job success or for training. Literally thousands of correlation coefficients indicating low relationships between occupational or academic performances of subjects and their scores on tests appeared in professional journals in the 1920's to the 1940's. Behind each ran the philosophical thread, ". . . everything which exists can be measured . . . and everything that can be measured exists." This was indeed empiricism run rampant!

John Dewey,¹³ writing in 1928, cautioned that the primary responsibility of school was the encouragement of diversity of performances rather than the uniformity that averaging of scores encourages. And he viewed with alarm the increasing efforts toward "scientific" measurement in schools. As statistical sophistication became more widespread other writers saw the dangers of the nomothetic or actuarial approach to the solution of problems of prediction. Recently Rothney and Roens pointed out that ". . . regardless of the general relationship between two variables expressed by a correlation coefficient, it is possible to find relationships within a particular counselee that run counter to the pattern indicated by the coefficient even to the degree of complete reversal of it; to find, within one person, the amount of correlation between characteristics which is common to the whole group; and even to find within one person, closer relationships than would be expected in view of the size and direction of the coefficients obtained from mass data."¹⁴

As the counseling movement grew and counselors clarified intents and purposes, it became clear that while the actuarial approach permitted fairly accurate predictions of group performances, the counselor was required to consider that fraction of the

¹³ John Dewey, "Progressive Education and the Science of Education," *Progressive Education*, August, 1928, 5:197-204.

¹⁴ Rothney and Roens, *op. cit.*, p. 19.

population for whom the prediction was not accurate just as much as he was concerned with the performances of the group as a whole. And with this recognition of his task, he became as much impressed with the optimum human development of each of his counselees as with the maximum utilization of manpower of any particular class of individuals. It became difficult to keep the proper perspective of dedicated interest in the individual counselee and the validation of clinical "hunches" about his future possibilities, when the generally accepted practice was one of assigning an index number to an individual and reading his predicted outcome from an actuarial table. While these practices seemed to work with a fair degree of success for such groups as the entering class of freshmen at a large university or training groups in the armed services, these procedures failed to predict the subsequent performance of many subjects on the job or in training.

That some low-scoring testees did far better on the job or in training than had been predicted from their scores, or that the ultimate performance of others who scored high on tests turned out to be disappointingly low, tended to make the sophisticated counselor wary of making performance estimates *for the individual* based on estimates of performance of groups. Careful search of many test manuals failed (and still fails) to disclose hints or suggestions as to actions of a counselor when confronted with this dilemma. Not completely satisfied with the dismissal by statisticians of this contradiction with statements about sampling errors or chance factors or probability estimates or maximum likelihood estimates, the concerned counselor began to challenge some commonly held psychometric concepts and to examine them with more care and detail. Detailed discussion of such matters will be found in later chapters.

TESTING IN MANPOWER UTILIZATION AND IN PERSONAL DEVELOPMENT

The increasing sophistication of some counselors in the use of

tests has by no means become universal. As suggested in the opening paragraph of this chapter, varying concepts of the role of testing and the use of test results still exist. This appears to be true not only in broad areas of application but also, to some extent, within the single area of counseling. Super, in a discussion of some of the differences in concepts underlying European guidance programs and those found in the United States, has shown the implications involved in the use of tests as tools for human development or as manpower utilization. His discussion of the French programs provides a good example of the "manpower utilization" point of view and the use and efficacy of tests in its implementation.

During part of our formal discussion the focus of the Committee's [French] interest was on the question of how we know what numbers of men and women will be needed each year in each occupation, how this information is applied at national, state, and local levels in the planning of educational facilities and programs, and, *to a much lesser degree, how it is applied to individuals* in planning an education and in choosing a field of work. The assumption was made that the information could be applied to the individual by testing his capacities and interests and finding out whether or not he qualified to enter the desired type of educational program. Obviously, if 100,000 secondary school teachers are needed, only the 100,000 best qualified candidates should be admitted to teachers colleges, and those falling below the cutting point should be diverted to other types of professional or occupational education. This type of application to the individual was more or less taken for granted. . . .¹⁵

This on the surface might appear to be a highly commendable and objective approach, except that it ignores the great limitations of tests as predictive devices, and worse, even if it recognizes the fallibility of tests, it relegates individual worth and integrity to a position of secondary consideration.

¹⁵ Donald E. Super, "Guidance: Manpower Utilization or Human Development," *Personnel and Guidance Journal*, September, 1954, 33:8-14.

Super's description of the French approach has been confirmed by Conant.

How sharp the difference . . . between our schools and those of [free] Europe! We use our schools to cultivate equality of opportunity. Europe uses its schools to keep wide the gap between those who will make a living by their hands and those who will make a living using their brain.

For children up to the age of 11 (in some cases 12 or 13), Europe provides good common schools. Then comes the tragic break. Children are sifted by tests and examinations to decide who will go on to secondary school and university and who will prepare for work. By this method, nine children out of ten are declared ineligible for further education and are shunted into training.¹⁶

The "manpower utilization" concept and its consequent implication about the use and high efficacy of tests is not, of course, peculiar to European guidance and educational programs. To a considerable extent, this viewpoint was basic in the use of tests by the military in this country during World War II. The use of tests for such purposes in the war under an emergency is understandable where screening of large numbers of men must be accomplished rapidly. The lack of time, the urgency of the situation, and the need for rapid and effective training demanded methods that could be applied to masses of men in the hope that the number of "successes" would exceed the results of assignment by lot. Here, of course, in the total job to be done—winning a war—personal wishes and aspirations were of secondary consideration.

To a considerable extent, a parallel to the military use of psychological tests is found in current industrial testing. Here again expediency dictates practice. Where the training of employees is expensive—\$5,000 to \$10,000 per man in some cases—the desire of management and stockholders to reduce failure of employees on the job is understandable. Increasing the number of successes

¹⁶ Staff, *The Teacher's Letter*, Washington, D.C.: Arthur C. Croft Publications, February 3, 1936.

over the number of failures shows up on the profit and loss statement, and even a slight improvement over chance selection in industry may justify the use of tests.

It has not been the intent of the authors to question the value of tests in the situations described, or in other applications outside the counseling in schools. The intent here has been to suggest that situations dictate the flexibility or rigidity of technical demands on the instruments used. As suggested earlier, the counseling situa-

TABLE 1. Factors Contributing to Counseling

Testing for Counseling (Secondary School)	Testing for Selection (Industry, Military, etc.)
Concern with <i>all</i> members of a particular situation, i.e., all students in a given high school regardless of range of performances and characteristics.	Concern with <i>limited</i> numbers of applicants for work within a specific organization, with some screening, perhaps, involved in the nature of job announcements and specifications.
Unique concern with one individual at a time. Counseling is an individual affair. Averages or percentages or success is of little comfort to those who are not successful.	Testing that improves over chance selection pays off in terms of production, the company's primary concern. Where there is a gain of even one successful employee over failures, testing may more than pay for itself.
Same obligation to all students. Individually, they are very much present and a working part of school organizations. Counselors cannot turn them away.	In a sense no obligation to any applicant and especially none to those not selected. Future or next steps of those rejected in a hiring situation require no further contact, or the formulation of alternate plans.
Students are going into a future the dimensions of which are not known. Counselor working with many variables and with many unpredictables.	Selection made into a defined situation—usually a specific job with dimensions established.
Concern with the individual for his own sake, his worth as an individual, his successes.	Not concerned with individual as such— <i>who</i> does the job not as important as getting the job done by anyone.
Many variables in persons and situations appear over a long period of time. Demands of society and differences in the definition of success by society do not permit many generally accepted definitions of success.	Selection testing can prove itself over the long range. It is effective if, over the long run, more successes than failures are picked.

tion has its unique demands, differing at times from those of industry. Its objectives are dictated, at least in part, by the philosophy of education that emphasizes human development. Some of the factors that contribute to making the school counseling setting different from that, for instance, of the military or industrial, are summarized in Table 1.

THE PLACE OF TESTS IN COUNSELING

In this volume the reader will be constantly reminded that his function is to serve individuals as individuals and that he is not employed primarily as a selector for industry or advanced educational institutions. It is implied that if he does the first task well the selective processes may be improved. It will also be noted throughout that the counselor will be working primarily with so-called normal cases in the setting of the American school rather than with pathological, "clinical," or "disturbed" cases in hospitals or with the applicant for employment in an industrial setting. It is implied that his effectiveness with normals and in the school situation may be improved by use of tests.

It is conceivable that much good counseling can be done without use of tests and, indeed, there is evidence that it was done many centuries before standardized tests were available. If all tests were currently eliminated from counseling it is unlikely that society would recognize the change for many years. Business would go on, schools would continue, and millions of young persons would be satisfied with their choices of training, occupation, marriage partners, and leisure-time activities. Research in education, guidance, and psychology has not clearly and conclusively demonstrated that the use of tests has increased the welfare or productivity of any significant numbers of persons despite the fact that millions of them are used annually. There is some evidence that they may be of assistance occasionally in answering the questions of persons who score at the extremes of test distributions but they do not

provide final answers. They may, for such persons, help in the working out of probabilities, odds, or best bets in general on the average, on the whole, and other things being equal (which they seldom are), and in doing so they may become valuable sources of supplementary information about students.

It seems likely that if tests are to be useful in counseling they will be so only insofar as they have been selected for use in answering *specific questions of particular* counselees. Since certain kinds of questions, such as those that refer to reading performances, are likely to be asked by many subjects, it may be desirable to administer a reading test to large groups of subjects at one time. Except in seeking answers to very commonly asked questions, however, it seems that testing is likely to be most useful when a plan of tailoring the testing program to individual cases is employed. It would, for example, be a considerable waste of time and money to give a so-called mechanical aptitude test to all the students in a general public high school because only a small sample of the counselees, their parents, their potential employers, or school personnel will or should raise questions about their mechanical performances. Those about whom such questions are raised may be tested individually or in small groups with, perhaps, some benefit.

The statements that appear above were designed to suggest to the reader that, with a few exceptions, testing for counseling is a *differential*, not a *mass* procedure. It is indicated that testing will be focused on particular individuals, that tests will be selected when there is a specific question to be answered. Tests will be used when other sources do not provide answers. "When in doubt, punt," says the football coach; and "When in doubt, lead trump," says the bridge player. Though neither is an infallible rule, the counselor may take a cue from them and decide that "When in doubt, test." When he does so he may sometimes score well and he may be able to assist counselees and others who are concerned with their welfare to answer questions (*in terms of probabilities*) that will help them to make wise choices.

In the following chapter we shall be concerned with the numbers and kinds of measuring instruments available and some of the problems involved in the selection of the most serviceable among them.

SUMMARY

In this chapter it has been suggested that psychological tests are used in great numbers each year in a variety of situations, each with its own rationale, and with varying implications for the lives of the persons involved. The nature of the individual places unique demands upon the application of psychological tests in counseling. The counselor must be prepared to help one student at a time with a wide range of problems and in the making of many decisions. In fulfilling this role he must turn to many sources of data and to a variety of techniques, no one of which is infallible. Psychological tests, as one of the sources of information about individuals, may assist the counselor in answering some of the questions with which he is faced. They are likely to be most useful where they are selected for use in answering *specific* questions about *particular* counselees and to the degree that it is recognized that testing in counseling is a *differential* and not a *mass* procedure.

DISCUSSION QUESTIONS AND EXERCISES

1. There is evidence that an increasing number of institutions of higher learning are turning to the use of psychological tests in admission procedures with a view to holding down enrollments. Such selection devices may result in some being accepted and failing and others being rejected who might have succeeded. What are the social implications of such a policy? How does such a policy relate to the concepts of human development? Manpower utilization? Who, if anyone, assumes the responsibility for each of the types of persons above and what could be done about the situation as described?
2. Assuming you, as a counselor, are asked one of the questions listed

on pages 7-8 (select one). What would you want to know about the individual, and what questions would you ask yourself about the situation before attempting to help the person to reach an answer or make the decision implied by the question?

3. The following request is the sort that is frequently received by guidance departments:

Dear Mr. Jones:

My son, age 16 and a Junior in high school, does not seem to know what he wants to do when he *graduates from high school* and I am concerned about it. I understand that there are tests to tell a youngster what he should do and I would appreciate it if you could arrange to have him take them.

Sincerely,

John Q. Smith

Prepare a reply to the letter. Keep in mind that the person making the request is not likely to understand professional jargon.

4. The following tables were recently drawn up from results of a questionnaire survey by a state department of education of testing practices in Grades 1 to 8 of the schools in sixty counties of one state. In order to get some uniformity in the interpretation of the items on the questionnaire, at least one conference was held with the person who was to answer it by a member of the staff of the state department of education. Comment on the purposes listed in Table 2 with respect to the relative rank of importance as indicated by the frequencies for each of the categories. If you were to rearrange them in the order in which you thought tests would be most useful *what rank would you give each of them? Would you eliminate any? If so, which and why? Would you recommend any changes in the kinds of resource persons (see Table 3) they would use? Why? What factors seem to have influenced the choice of tests listed in Table 4? Using the general concepts suggested under "Testing in Manpower Utilization and in Personal Development," how would you classify each of the stated purposes?*

Measurement for Guidance

TABLE 2. Purposes of the County Testing Program

Purposes	Number of Counties
Measuring general achievement of individual pupils	45
Diagnosing pupils' needs and abilities	38
Grouping pupils for instruction	35
Determining mental ability	30
Determining basis for remedial instruction	23
Securing data for guidance purposes	20
Aiding teachers in self-evaluation	13
Deciding whether curriculum objectives were met	9
Deciding on promotion of pupils	9
Evaluating achievement on a county-wide basis	8
Determining readiness	8
Helping supervisors in work with teachers and/or pupils	7
Confirming teachers judgments of pupil progress	6
Interpreting pupil progress to parents	5
Providing a basis for grading	5
Developing pupil test "awareness"	4
Making general comparisons	3
Providing defense against public criticism	1

TABLE 3. Resource Persons Used by Counties in Planning of Testing Programs

Resource Persons	Number of Counties
State department of education personnel	28
Testing company representatives	23
College or university personnel	19
School personnel from other counties	10
Principals of local schools	9
Psychologists from guidance or welfare departments	6
State employment service	2
Textbook publishing company representatives	1

TABLE 4. Names and Frequencies of Mental Ability Tests Used in County Testing Programs

Names of Tests	Number of Counties
California Test of Mental Maturity	30
Otis Quick-Scoring Test of Mental Ability	12
Kuhlman-Finch	5
Pintner-Cunningham	3
Hennon-Nelson Test of Mental Ability	2
Davis-Eells	1
Chicago Non-Verbal	1
Kuhlman-Anderson	1

5. The following statement is from a pamphlet related to the use of tests in employment situations:

A well-chosen psychological test may be thought of as a written interview. The group test possesses the added advantage of lending itself to mass-production methods in the obtaining of information needed in the placement of new-hires and in appraising the desirability of transferring an unsuccessful or disgruntled employee already on the job.

To be a valuable tool in the hands of the employment interviewer, a test must tell what would be found out when the supervisors' ratings start coming in. Tests give this information in fewer minutes by the clock, sooner by the calendar, and at a lower per-employee cost. The group test is especially valuable in that it permits the simultaneous "interviewing" of 25 to 100 applicants, thus enormously reducing delays in getting new-hires on the job.¹⁷

What assumptions regarding the efficacy of testing are apparent in the statement? How does the viewpoint presented above contrast with that of individual counseling? Is the argument of time-saving tenable even in employee selection? Would you agree that "a well-chosen psychological test may be thought of as a written interview"?

REFERENCES

- Cronbach, Lee J. *Essentials of Psychological Testing*. New York: Harper, 1949, Chapter 11.
- Darley, John G., and Anderson, Gordon V. "The Functions of Measurement in Counseling." In E. F. Lindquist (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 68-84.
- Doppelt, Jerome E., and Bennett, George K. "Reducing the Cost of Training Satisfactory Workers by Using Tests." *Personnel Psychology*, April, 1953, 6:1-8.
- Fortune*. "The Tests of Management." *Fortune*, July, 1950, 42:92-96.

¹⁷ Floyd Ruch. *How to Use Employment Tests*. Employment Testing Bulletin No. 1. Los Angeles: California Test Bureau, 1944, p. 3.

- Jones, Arthur J. *Principles of Guidance*. New York: McGraw-Hill, 1945, Chapter 2.
- Mathewson, Robert H. *Guidance Policy and Practice*. New York: Harper, 1955.
- Rothney, John W. M., and Roens, Bert A. *Counseling the Individual Student*. New York: Dryden Press, 1949.
- Rothney, John W. M., and Roens, Bert A. *Guidance of American Youth*. Cambridge: Harvard University Press, 1950.
- Super, Donald E. *Appraising Vocational Fitness*. New York: Harper, 1949, Chapter 2.
- Super, Donald E. "Guidance: Manpower Utilization or Human Development." *Personnel and Guidance Journal*, September, 1954, 33:8-14.
- Thorndike, Robert L. *Personnel Selection*. New York: Wiley, 1949.
- Warters, Jane. *High School Personnel Work Today*. New York: McGraw-Hill, 1956. Chapter 1.
- Whyte, William H., Jr. "The Fallacy of Personality Testing." *Fortune*, September, 1954, 50:117-119.
- Wolfe, Dael. *America's Resources of Specialized Talent*. Report of the Commission on Human Resources and Advanced Training. New York: Harper, 1954.

CHAPTER II

Varieties and Sources of Tests

Tests that purport to provide the answers to practically all the questions that counselors and their counselees are likely to raise have been published. Examinations of publishers' catalogues reveal hundreds of tests with titles inferring that measurement of aptitudes for, and performance in, common academic fields and vocational areas can be done successfully.¹ The bewildering array of titles of tests with similar content and designated purpose requires the counselor to develop standards to guide him in the selection of those that he will use.

TYPES OF TESTS

CLASSIFICATION AND EMPHASIS

In the following chapters some criteria to aid in selecting tests will be considered. Before those criteria are examined it will be necessary to appraise the common types of tests in terms of the

¹ Measures of interest and personality are not considered here since the instruments are not tests in the usual sense of the word. They are questionnaires and as such will be treated separately in Chapter VIII.

- Jones, Arthur J. *Principles of Guidance*. New York: McGraw-Hill, 1945, Chapter 2.
- Mathewson, Robert H. *Guidance Policy and Practice*. New York: Harper, 1955.
- Rothney, John W. M., and Roens, Bert A. *Counseling the Individual Student*. New York: Dryden Press, 1949.
- Rothney, John W. M., and Roens, Bert A. *Guidance of American Youth*. Cambridge: Harvard University Press, 1950.
- Super, Donald E. *Appraising Vocational Fitness*. New York: Harper, 1949, Chapter 2.
- Super, Donald E. "Guidance: Manpower Utilization or Human Development." *Personnel and Guidance Journal*, September, 1954, 33:8-14.
- Thorndike, Robert L. *Personnel Selection*. New York: Wiley, 1949.
- Warters, Jane. *High School Personnel Work Today*. New York: McGraw-Hill, 1956. Chapter 1.
- Whyte, William H., Jr. "The Fallacy of Personality Testing." *Fortune*, September, 1954, 50:117-119.
- Wolfe, Dael. *America's Resources of Specialized Talent*. Report of the Commission on Human Resources and Advanced Training. New York: Harper, 1954.

CHAPTER II

Varieties and Sources of Tests

Tests that purport to provide the answers to practically all the questions that counselors and their counselees are likely to raise have been published. Examinations of publishers' catalogues reveal hundreds of tests with titles inferring that measurement of aptitudes for, and performance in, common academic fields and vocational areas can be done successfully.¹ The bewildering array of titles of tests with similar content and designated purpose requires the counselor to develop standards to guide him in the selection of those that he will use.

TYPES OF TESTS

CLASSIFICATION AND EMPHASIS

In the following chapters some criteria to aid in selecting tests will be considered. Before those criteria are examined it will be necessary to appraise the common types of tests in terms of the

¹ Measures of interest and personality are not considered here since the instruments are not tests in the usual sense of the word. They are questionnaires and as such will be treated separately in Chapter VIII.

methods of administering them, the materials used, the performances demanded of the subject, and their purported use.

With respect to *administration of the tests*, they may be given to *groups* or to *individuals*.

Tests may produce one score or a series of scores. Those that produce one score may be called *unit* tests and those that provide several may be described as a *battery*.

Tests may consist of *miniatures* of larger tasks or they may contain items from which a *trait* is inferred.

With respect to the *materials used* in a test, the items may be in *verbal*, *nonverbal*, or *apparatus* form.

The *time factor* in testing may be represented by tests in which a definite *time limit* is prescribed or those in which a *work-limit* procedure is used. In the latter case the subject is given as much time as he needs to complete the items.

The performances demanded of the subject may test his *speed* in responding, his *power* within a comparatively narrow area, or his *range* of coverage in several areas.

The *extent of generalization* about the subject's test scores may range from a *specific* report about a particular area such as spelling or a *general* indication of as broad a factor as general scholastic aptitude.

The area measured by a particular test may be described by the author as *scholastic*, *mechanical*, *scientific*, *numerical*, *cultural*, *clerical*, and so forth.

The *function* measured may be variously labeled as *achievement*, *aptitude*, *ability*, *information*, *proficiency*, etc.

The number of combinations of the categories noted above are many and they increase as new tests appear and new titles are given to tests that cover the areas formerly covered by older tests. The schematic diagram of Figure 1 has been drawn up to indicate some of the possible combinations. Using the items in the diagram, it is possible to classify many of the common tests. Thus the Stanford-Binet may be described as an *individual*, *unit*, *trait*, *verbal*

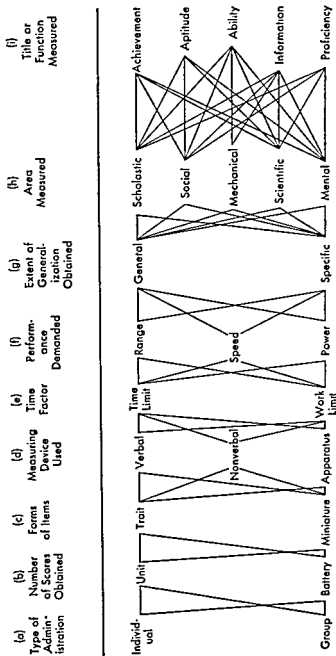


FIGURE 1. A Schematic Presentation of Common Types of Tests.

and *nonverbal, time limit, and power test of general mental ability*. The Stanford Achievement Test, by contrast, may be classified as a *group, battery, trait, verbal and nonverbal, time limit, range test of specific scholastic achievements*.

IMPLICATIONS FOR SELECTION

The analysis above was presented to indicate the variety of tests from which a counselor must choose. Thus if a counselee were to raise a question about his fitness to undertake an apprenticeship in a mechanical field and if the counselor decided to use tests to assist him in answering that question, he would have to make decisions about such problems as these:

1. Should I use a group or an individual test?
2. Should I use a test the materials of which duplicate in miniature the actual mechanical tasks he will be required to perform in the apprenticeship, or should I use a test that purports to measure general mechanical aptitude?
3. Should I use a battery of tests (spatial relations, speed of manipulation, mechanical reasoning), or will a test that provides a single score be adequate?
4. Should I use tests that require the subject to perform on a piece of mechanical apparatus? Can I get the answer that I seek by having the subject write his responses? Or shall I require him to manipulate symbols as in a spatial relations test?
5. Should I let him take a test that permits him to use as much time as he needs to complete the tasks set, or should I choose a test with definite time limits?
6. Should I require him to perform in several areas, or should I see how far he can go in selected specified areas?
7. Should I be concerned with scores in many mechanical areas that can be averaged to give a general mechanical score, or should I test him intensively only in the special part of the mechanical field in which he has expressed interest?

8. Should I ignore his performances in scholastic or other areas and concentrate entirely on the mechanical field?

9. Should I be concerned primarily with the information he has gained in the mechanical area, with the kind of reasoning he does when presented with mechanical problems, or should I be concerned with his proficiency in performing certain mechanical tasks only?

When the counselor has answered these questions he may begin looking for tests to fit the specifications set up by the answers to his questions. But in the process of answering them he will have been forced to consider many of the controversies about the value of various testing devices. Some of them are presented in the following paragraphs.

GROUP VS. INDIVIDUAL TESTS

In deciding whether he should use a group or an individual test he must weigh the advantages of saving time by testing several individuals simultaneously against the loss of the opportunity to observe a particular counselee closely while he is at work. He must consider the possibility that his subject might not put forth maximum effort in a less closely supervised group-testing situation and the chance that he might be disturbed by the presence of others. In making the choice between individual and group tests he must study the reported difference in the yield of group and individual tests and try to determine whether the differences are worth the additional time the individual test requires.

SINGLE SCORE VS. BATTERY TESTS

When a counselor is trying to decide whether he will use a test that provides a single score or a battery of tests yielding several scores he will be forced again to consider the factors of time and yield. Certain tests labeled as tests of ability or aptitude or reason-

ing in mechanical areas purport to give, in a single score, a measure of what the test has been labeled. This practice is continued despite the fact that investigators have shown rather conclusively that it would be better to talk about mechanical *abilities* rather than mechanical *ability*. Such evidence suggests that he will be forced to use a battery of tests rather than a single test. Since, however, the battery will be more expensive in time and money the counselor must weigh relative contributions of unit tests and batteries in his decision.

MINIATURE VS. TRAIT TESTS

As he proceeds in the selection of his tests, the counselor may have to make a choice between miniature and trait tests. For a counselee who is considering a career as a machinist the counselor may, for example, use a two-hand coördination test in a setup similar to a lathe or he may ask a set of questions designed to determine whether his subject has become familiar with the principles involved in lathes and similar machines. In making this choice he will be faced with the problem of weighing the cost of a potential increase in yield obtained by using one of the methods. He must also consider whether his miniature can ever fully represent the whole and whether or not he wants to approach his problem from as narrow an approach as the use of miniatures may require.

RECOGNITION VS. DEMONSTRATION TESTS

This next problem relates closely to the one in the paragraph above. Shall he use a picture-type test in which the subject is asked to show his facility in interpreting pictorial representations of certain mechanical operations? Shall he use the spatial relations type of test in which his counselee is required to recognize certain designs when they are turned in different directions or to recognize

the kinds of designs that may be produced from several component parts? Or shall he use a piece of apparatus, miniature or otherwise, on which his counselee will *demonstrate his facility*? He will need to examine the research concerning the relative contribution of each of these kinds of tests to determine which of them, separately or in combination, best predict the future performances of subjects in the mechanical field. He will find little evidence of the kind he seeks.

As he continues to seek answers to the remaining questions he will, of course, find some overlap with those to which he has sought answers previously. In trying to decide between work-limit and time-limit tests the merits and limitations of each of these testing methods must be appraised. He will find advocates of both methods and his examination of research on this problem will probably not result in clear-cut answers to his questions.

GENERALITY VS. SPECIFICITY OF TEST COVERAGE

During the time that the counselor has been seeking answers to the above questions he will have been concerned with the generality or specificity of the coverage of the area that he wishes to cover. He may wish to know whether the counselee is likely to succeed in any of several mechanical tasks or, in rare cases, he may be concerned with both. In general he will discover that the tendency in testing is to turn away from global and to seek specific scores. He will also find that, in industry, the trend is toward selecting men who are generally proficient and who may be taught a specialty after they have been employed.

TEST TITLES VS. TEST CONTENT

Finally, the counselor will look at all possible tests in terms of what their labels indicate they purport to measure. He will find tests that are labeled as *measures of mechanical reasoning*, me-

completion of analogies may be a part of a verbal aptitude test in one case but part of a general mental ability test on the other.

The exercise above should have revealed to the reader that he must be concerned with test items rather than with test titles. If he is still skeptical he should try still another exercise. He should arrange to have someone take a sampling of various tests and fold their covers back so that the reader cannot see the title. The reader may then study the items and guess what the title might be. He will find little correlation between the titles guessed and the titles given. The title on the cover of a test gives little indication of its content.

SOURCES OF TEST MATERIALS

Now that the counselor has noted the various kinds of tests available and has seen some of the problems he will meet in making choices among them, he may inquire about the sources of materials. In the following pages some samples of basic sources are called to the attention of the counselor. After he has considered them he may then turn to references in which the names of publishers of specific tests are given.^{2, 3}

NONPROFIT TESTING AGENCIES

The Educational Testing Service of 20 Nassau Street, Princeton, New Jersey, is sponsored by the American Council on Education. The council is composed of national education associations, universities, colleges, technological schools, private secondary schools, city school systems, state departments of education, and other educational groups. It is a center of coöperation and coördination whose influence has been apparent in the shaping of American

² O. K. Buros, *The Fourth Mental Measurements Yearbook*. Highland Park, N.J.: Gryphon Press, 1953, pp. 1100-1106.

³ C. C. Ross, *Measurement in Today's Schools*. (Revised by J. C. Stanley.) New York: Prentice-Hall, Inc., 1954, pp. 464-465.

educational policies and the formation of educational practices for nearly forty years. Its committees are composed of famous persons in American education including presidents of universities and colleges of high standing and, ex officio, the United States Commissioner of Education.

Within recent years the Educational Testing Service has, under the sponsorship of the Council, carried on the testing activities formerly done separately by the College Entrance Examination Board, the Carnegie Foundation for the Advancement of Teaching, and the Coöperative Test Service. It was organized to serve education by developing new areas in which tests were needed, by constructing and administering testing programs for various educational and government purposes, by conducting research for the purpose of advancing test theory and practice, and by providing advisory services to schools and colleges. It has provided testing programs for many government and private scholarship programs and has prepared batteries of tests that are used as admission and evaluation instruments for many colleges and graduate schools.*

The counselor should become acquainted with the Educational Testing Service. Its nonprofit basis and its high quality of research and offerings demand that one consider its services before selection of tests for any purpose is made.

The Educational Records Bureau, 21 Audubon Avenue, New York 32, New York, also sponsored by the American Council on Education, has maintained high standards of testing for large numbers of private and public schools on a nonprofit basis. It has maintained a test research staff and its publications contain many reports of tryouts of commercially published tests.⁵ Although the recent advocacy by the Bureau of some questionable interest inventories seems contrary to the high standards it has maintained, its offerings are generally excellent. Specific information about the

* The Annual Reports to the Board of Trustees describe the activities of ETS in detail. Literature concerning the offerings may be obtained on request.

⁵ Bulletins of the Educational Records Bureau published annually.

Bureau's services may be obtained at the address given above. No counselor can afford to be unfamiliar with its offerings.

COMMERCIAL TEST PUBLISHERS

The test publishers who sell tests for profit may be grouped into four categories. Since there are many publishers in each of three of the categories, only samples of names are presented. Complete lists may be seen in *Buros' Mental Measurement Yearbooks* mentioned previously.

The first category includes commercial test publishers whose primary purpose is the sale and distribution of tests. These organizations are not sponsored nor supervised by professional organizations, regardless of the titles they have chosen for their business. They are maintained and continue to operate on the basis of profits obtained from the sale of tests. Some of the organizations in this category are:

Science Research Associates. 57 West Grand Avenue, Chicago 10, Illinois.

California Test Bureau. Head office, 5916 Hollywood Avenue, Los Angeles 28, California.

Public School Publishing Co. Bloomington, Illinois.

Educational Test Bureau. 720 Washington Ave. S.E., Minneapolis, Minnesota.

Committee on Diagnostic Reading Tests. Kingscote, Apt. 39, 419 West 119th Street, New York 27, New York.

A second category includes several commercial companies whose primary concern is the publishing of books and who also distribute tests and related materials. Most of these organizations are members of The American Textbook Publishers' Institute and such membership offers a preliminary screening for the counselor who is seeking quality products. Many of these publishers have established enviable reputations in their fields over a long period of

time and are not likely, if they can avoid it, to jeopardize that reputation by offering unsatisfactory tests for sale. Unfortunately it has not always been possible to avoid this situation. The counselor should, however, *examine the offerings* of such companies as the following when he begins to select tests:

Houghton Mifflin Co. 2 Park Street, Boston 7, Massachusetts.
World Book Co. Yonkers, New York.

The third group includes university presses that publish a number of psychological tests and diagnostic devices. They are, of course, under the control of universities of high international repute and their editorial responsibility is at a high level. Again, some of these organizations produce tests and inventories that may seem of doubtful value to the counselor. Acceptance of the high ethical standards of the publisher does not imply complete acceptance of their test materials nor the theories on which they are based. *Examination of the offerings of such organizations as those noted below, and indeed of any of the university presses, is recommended to counselors.*

Stanford University Press. Stanford, California.

Harvard University Press. Cambridge 38, Massachusetts.

Bureau of Publications, Teachers College. Columbia University, New York 27, New York.

A fourth category is added because one major commercial source does not clearly fall in the first three. The Psychological Corporation, 304 E. 45th Street, New York 17, New York, is a special case. It was organized some 25 years ago to provide *instruments and techniques* developed by the psychological profession. Its board of directors, officers, and staff consists of members of the American Psychological Association and the ownership of its stock is restricted by its charter to members of that association. It is composed of divisions devoted to *market and social research, industrial, clinical, and professional examinations and tests.* Its Test

Service Bulletins,* which may be obtained without cost, are models of a professional service that can be used by counselors. The direction of The Psychological Corporation's efforts by members of the American Psychological Association suggests that high ethical standards are maintained. It is not suggested, however, that the counselor must agree with their basic thinking on measurement or accept without question all the materials they produce.

TESTS USED BY AGENCIES OF GOVERNMENT

Under this heading the counselor will note such tests as the Selective Service College Qualification Test, the General Aptitude Test Battery of the United States Employment Service, and perhaps the state-wide testing services. In the case of the first of these he will not have any voice in the construction or administration of the tests but counselees may consult him about times at which the tests are to be taken and about the interpretations and use of scores.

The General Aptitude Test Battery, consisting of 12 subtests and designed to measure aptitudes said to be necessary in certain occupational areas, is administered and scored by State Employment Services. In many schools, representatives of the agency seek to administer the battery to high school seniors who plan to enter employment immediately after graduation. They will interpret the scores to such volunteers and make them available to school counselors in specific cases.

The counselor who plans to use such services must first seek answers to several questions and make certain decisions about testing policies. He may find, for example, a tendency for economy-minded and antiguidance school personnel to proclaim that they need not be concerned with guidance since the employment services will do what is necessary. He must ask whether these tests, coming in the

* Test Service Bulletins. Issued irregularly. They may be obtained free by writing to The Psychological Corporation, 304 E. 45th Street, New York 17, New York.

senior year, are administered too late to be of much value. He must decide if he wants a nonschool agency to determine the kinds of tests that would be most suitable for his counselees. He must consider whether the coöperation of the school with employment services and employers in the use of tests will not result in better relationships among all three agencies. And, finally, he must decide whether tests designed for use in selection may be useful in counseling.

On this latter point the evidence is not clear. One author⁷ has pointed out the following difficulties in use of the test: There is no intelligible guidebook for the interpretation of test results; national norms for high school seniors by sex that would permit comparison of a senior's score with a clearly defined group to which he belongs are not available; percentile norms of young high school graduates in various occupations are not provided; reliability data on the test when it is used with high school students and probable errors of measurement for various groups are not offered; data on predictive validity in terms of comparisons of performances of high school seniors with well-defined criteria of job success after employment are not given. At least one follow-up study of the test battery⁸ has shown no significant differences in job satisfaction of a group of young workers who had taken the tests and received interpretations of their scores while in high school and a matched group that had not.

In one sense the problems and questions raised above reflect those that are commonly met when the counselor becomes involved in testing programs in which he cannot select his own tests, set up the rules for the testing and interpreting the scores to his counselees, and tailor the measurement program to particular individuals. There are, of course, some compensations. He is, in general,

⁷ T. E. Christenson. "Helping Students Enter Industry." *Vocational Guidance Quarterly*, Autumn, 1954, pp. 24-26.

⁸ Carl Traeger. "Effectiveness of the United States Employment Service General Aptitude Test Battery in Employment Counseling of High School Seniors." Unpublished Ph.D. Thesis, Madison: University of Wisconsin, 1955.

likely to get better tests at less cost than he may be able to procure from commercial test publishers. In state or regional programs,⁹ usually centered at state universities, he is likely to get norms on subjects that are similar to his counselees, and he may have the advantage of participating in a professional approach to test development, study of fresh materials, and examination of new techniques. No counselor can overlook the offerings of such agencies when he plans his measurement program.

LOCALLY CONSTRUCTED TESTS

Counselors who are employed in educational institutions where there are large enough enrollments to permit development of satisfactory norms may find it to their advantage to use locally devised tests. The obvious limitations presented by using local materials for an essentially mobile population do, however, limit the value of this source except in unusual circumstances. As indicated later in this volume, the use of locally constructed achievement tests may be satisfactory when current, but not necessarily predictive, evidence about a student's performance is needed.

CHECKING ON THE TEST PUBLISHER

Regardless of the sources to which a counselor goes to get his tests, he should apply certain criteria in judging their merits. Application of these criteria will be illustrated for particular tests in Chapters IV and V. The following six criteria may be listed at this point.

1. Have the publishers done enough research to demonstrate that the merits they claim for the test are valid?
2. Do the publishers limit the claims for the value of their tests to what can be demonstrated by research and usage? (This point

⁹ E. F. Lindquist, "Nationally Coordinated Regional Testing Programs in High School," *New Directions for Measurement and Guidance*, American Council on Education, Series I, No. 20, 1944, pp. 87-103.

is particularly important to counselors. Many authors of tests state that they may be used for educational and vocational guidance of *individuals* but provide data that refer only to use with *groups*.)

3. Do the publishers continue to assume responsibility for a test after its publication? Do they provide revisions and improvements when their need is indicated?

4. Do publishers actually limit the sale or distribution of tests to qualified users? Statements of policies with respect to this matter have been published in *The American Psychologist*, August, 1946, I, No. 8, pp. 353-357.

5. Do publishers provide professional services to test users? Some companies give titles to their salesmen which imply that they are *consultants*. *Guidance personnel should consider the definition of the word "consultant" before they seek counsel on the purchase of tests.*

6. Is the publishers' advertising ethical and dignified? In professional fields there is common agreement that advertising must meet such criteria. The counselor will look with as much suspicion on test publishers who resort to high-pressure advertising of testing programs to meet all his needs in a neat package as he would on the advertising of a patent medicine that is supposed to cure all his physical ills.

SUMMARY

In this chapter some of the problems of selecting a test from a wide range of types and forms have been presented. Tests have been classified with respect to *methods of administration*, the items they contain, the performances demanded of the subjects, the scores they produce, and their purported use. It has been shown that selection of the test or tests that might be most useful from among the many available is a complicated task. It requires examination of much research and the weighing of many kinds of conflicting evidence. It has been demonstrated that test labels are

generally misleading and that the counselor must be concerned with more than a test title. Sources of tests have also been indicated. In the following chapter the processes involved in the selection of a test, regardless of the title it carries, will be considered.

DISCUSSION QUESTIONS AND EXERCISES

1. Select several tests with essentially the same title or purportedly measuring such things as mechanical aptitude and classify each according to the descriptive terms suggested in this chapter. How do they compare? How would you account for the differences of approach? Which of those you have analyzed do you think represents the soundest approach, considering the nature of the area to be tested and the stated purposes of the tests? Why?
2. Using the tests selected for the exercise above, examine them for types of items used, e.g., pulleys, tool identification, spatial relations, etc. How do the tests compare in terms of types of items used? What types of items, if any, are common to each of the tests? What types of items are peculiar to a specific test? Are these consistent with the stated purposes of the test? How would you account for the differences? Is there evidence in the test manuals to suggest that the tests are equally effective in spite of the differences in test items? To what extent do the test authors defend the inclusion of their items?
3. As a class project, examine as many tests as you can from your specimen library and compile a list of traits, skills, achievement areas, personal characteristics, and so forth. Place this list in the right-hand column of a table such as Table 5. Provide a further column for each of the major test publishers. On the basis of data obtained from their catalogues and other publications, place a check mark in the appropriate line and column whenever a publisher offers a test or subtest score appropriate for use at the high school level which corresponds to the trait, skill, etc. listed in the first column. The partial table, presented as an example, suggests that publishing companies A, C, D, E, G, and H all offer a test designed to measure mechanical aptitude.

TABLE 5. Survey of Special Tests and Publishers

Trait, Skill, Achievement Area, etc.	Test Publishing Company							
	A	B	C	D	E	F	G	H
Mechanical aptitude	x		x	x	x		x	x
Neurotic tendency		x				x		x
Clerical aptitude	x		x	x	x		x	x
Reading speed		x		x		x	x	x

Can you think of any areas of performance or behavior that are not represented by the list of traits, etc., presented in the first column of your table? What areas of measurement appear most often among the test publishers? Are there any areas of measurement represented on which you believe it would not be desirable to have performance data? Would you regard it as desirable to give a student an extended battery of tests representing these areas listed in the first column? Why or why not? What publishers appear to provide the widest range of measurement? How would you account for the duplication in the offerings among the publishers?

4. A description of the job of cabinetmaker might include the following statements regarding the nature of the work:

Studies work orders, drawings, blueprints or other specifications; measures with such instruments as calipers, scale, square; may use hand tools such as rip saw, rabbet plane, files, hand plane; uses such woodworking machines as circular rip saw, tenoner; turns round parts to desired diameter on a lathe; forms such wood joints as butt joint, miter joint, mortise and tenon, and lap joints; checks vertical and horizontal trueness with carpenter's level; may apply oil, stain, or polish to complete articles; installs hardware such as hinges; estimates job costs; makes sketches or drawings of work to be done.¹⁰

¹⁰ Adapted from "Job Description for Cabinetmaker." *Occupational Guide* series, Occupational Analysis and Industrial Services Division, U.S. Employment Service,

For what kind of evidence would you look if you were counseling with a student who planned to enter training for this job? What types of tests would provide the best measurement in the areas suggested by the statements? Can you find examples of specific tests that would provide appropriate measurement? What evidence of the appropriateness is offered in the manuals? Do any of the manuals specifically mention the prediction of success in cabinetmaking or related occupations?

5. Evaluate one test-publishing company in terms of the six criteria presented on pages 42-43. Use the publisher's catalogue, copies of test manuals, advertising materials, and any other available sources of information. You may wish to compare your findings with others in the class who have selected the same publisher and with those who have selected other publishers. Do the publishers reviewed rate about equally well? Does there appear to be any particular area of weakness? Which of the criteria would you regard as most important? How do the publishers reviewed by the class rank on the basis of this one criterion?

TESTS FROM WHICH THE ITEMS ON PAGES 34-35 WERE TAKEN.

1. Stanford-Binet *Intelligence* Test
2. Progressive *Achievement* Test
3. Numerical *Ability* section of the Differential *Aptitude* Tests
4. Progressive *Achievement* Test
5. Number section of the tests of primary *Mental Abilities*
6. Language usage section of the Differential *Aptitude* Tests
7. Stanford-Binet *Intelligence* Test
8. Progressive *Achievement* Test
9. Language usage section of the Differential *Aptitude* Tests
10. Henmon-Nelson Test of *Mental Ability*
11. Verbal reasoning section of the Differential *Aptitude* Tests
12. Mechanical reasoning section of the Differential *Aptitude* Tests
13. McQuarrie Test of Mechanical *Ability*
14. Henmon-Nelson Test of *Mental Ability*
15. Bennett Stenographic *Aptitude* Test

REFERENCES

- American Educational Research Association. "Psychological Tests and Their Uses." *Review of Educational Research*, February, 1947, 17.
- American Educational Research Association. *Review of Educational Research*, February, 1953, 23.
- American Psychological Association. *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Joint Committee of the American Psychological Association, American Educational Research Association, and National Council on Measurements used in Education. Supplement to *Psychological Bulletin*, March, 1954, 51. Washington, D.C.: American Psychological Association, 1954.
- Anastasi, Anne. *Psychological Testing*. New York: Macmillan, 1954.
- Buros, O. K. *The Fourth Mental Measurements Yearbook*. Highland Park, N.J.: Gryphon Press, 1953.
- Cronbach, Lee J. *Essentials of Psychological Testing*. New York: Harper, 1949.
- Freeman, Frank S. *Theory and Practice of Psychological Testing*. Revised Edition. New York: Henry Holt, 1955.
- Goodenough, Florence. *Mental Testing*. New York: Rinehart, 1949.
- Hildreth, Gertrude H. *A Bibliography of Mental Tests and Rating Scales*. 1945. Supplement. New York: The Psychological Corporation, 1946.
- Kirk, Barbara A. "Test Distributors and Our Needs." *Occupations*, January, 1951, 29:257-259.
- Laitin, Yale J. "Why Publishers' Representatives Were Born." *Occupations*, January, 1951, 29:260-263.
- National Education Association. *Technical Recommendations for Achievement Tests*. Report of Committees on Test Standards of the American Educational Research Association and the National Council on Measurements Used in Education. Washington, D.C.: National Education Association, 1955.
- Super, Donald E. *Appraising Vocational Fitness*. New York: Harper, 1949.
- Traxler, Arthur E. *Techniques of Guidance*. Revised Edition. New York: Harper, 1957.

CHAPTER III

Criteria of Test Selection

The publication in 1954 by the American Psychological Association of its bulletin on technical recommendations¹ for use of psychological tests and diagnostic procedures was an important event in the field of measurement. The very fact that the publishers thought that the recommendations were necessary implied, and in some cases the comments suggested specifically, that there had been some serious shortcomings in tests and in reports about them. The bulletin reviewed common weaknesses in descriptions of tests and suggested many ways in which the shortcomings might be avoided. Thorough study of this bulletin is an obligation of all those who produce, distribute, and use tests but, as the authors have pointed out, the publication of adequate information about them does not guarantee that tests will be used wisely or well. In this chapter, the reader will find suggestions for better selection of testing materials, and he will also be introduced to some of the problems that counselors meet in their use and interpretation.

¹ Technical Recommendations for Psychological Tests and Diagnostic Techniques. *Psychological Bulletin Supplement*, March, 1954, 51:1-38.

VALIDITY

One of the interesting phenomena in the field of education has been the widespread use of tests despite the fact that most of their manuals contained no adequate evidence that the tests could accomplish what they purported to do. When one refers to the question whether a test does what its authors claim that it can do (assess mechanical aptitude, determine readiness for reading, predict success in college, etc.) the term *validity* is used. In a sense the choice of this generally useful word in a technical sense is most unfortunate. A valid signature on a check may actually be the signature of the person but it does not guarantee that the check will be backed by sufficient funds. The name on a bottle of medicine may be the name of the product given to it by its makers but the medicine may not actually do what the manufacturers claim for it. It is conceivable that a test may be a valid measure of, say, mechanical reasoning, but it may fail to accomplish what the authors claim for it, that it will select the persons who are most likely to succeed in mechanical training or occupations. The counselor must always seek answers to such questions as: Valid for what? Valid for whom? and, Valid under what *circumstances*? And he will examine the evidence about the validity of a test with such questions in mind.

Until recently the general term "validity" has been used without any qualifying adjective. It is now becoming more common, after the urging of the committee that published the technical recommendations for use of psychological tests, to add an adjective before the word to indicate the kinds of evidence that are offered to substantiate the claim of validity. Currently the adjectives most commonly used are content, predictive, concurrent, and construct. Brief descriptions of the conditions under which counselors will be concerned with these four kinds of validity and of the ways in which test builders procure data about them are presented below.

CONTENT VALIDITY

A teacher may want to know the current level at which a student can perform on test materials drawn from subject fields in which he has been given instruction. Before he uses a test for this purpose he will need assurance that the items cover completely, or provide adequate samples of, the subject matter about which the conclusions are to be drawn. Thus if there is incontrovertible evidence that the items cover all the material that instructors have tried to teach their students to use in the way they are trying to get them to use it (and the school board members and parents in the community agree that the materials and the methods of using them are what they want for their children) the test will have content validity. If the test has been reduced in length by using samples of materials covered it will be necessary to have evidence that the sample is qualitatively representative and numerically adequate. If such evidence is presented, the content validity will be satisfactory and, providing other criteria of test construction and administration to be described later have been met, the scores may be used with confidence as a measure of a student's current performance in the area in which he was tested.

PREDICTIVE VALIDITY

The last sentence in the previous paragraph should be read again and particular attention should be paid to the word *current*. If tests with high content validity were available, and current performances always predicted future performances without substantial error, counselors would find themselves well equipped with tools to aid their clients in the choice of future academic and vocational activities. There is a great deal of evidence, however, which indicates that current tests of performances do not do so. It will

be necessary, if the counselor is to attempt prediction² of future performances of a counselee by means of test scores, to have some evidence of the *predictive validity* (in a sense the forecasting efficiency) of the instruments. If the tests are designed to measure mechanical aptitude, and if a youth's scores on them are to be used as one of the sources of evidence about likelihood of success in a mechanical occupation at some time subsequent to that at which he took the tests, there must be some evidence of the relationships between scores on the test and future performances in mechanical work. If so-called intelligence or scholastic aptitude tests are to be used as partial indicators of later achievement in college there must be some evidence that the test scores do forecast degrees of success in college.

When a counselor looks for *predictive validity* in a test he is not necessarily concerned with the content of the test or the form of the items, although it should be noted that, for successful interpretation of tests to subjects, it seems desirable to have some content validity. It is difficult, for example, to get a parent who is a mechanic to accept, as evidence of his son's fitness for mechanical work, the scores on a mechanical reasoning test that contains picture items of race tracks, billiard tables, and children on swings, as predictively valid as these items may be. It is theoretically possible that the speed with which a candidate for air corps pilot training ran a mile might predict later success in flight training, that achievement in spelling might predict future performance in accounting, or that skill in translating artificial language might predict eventual success in mathematics. And if there was evidence that they did so they might well be used as tests with high predictive validity in those areas. The point here is that there must be evidence of some relationship between scores on tests and later

² The argument about whether or not counselors should predict need not be considered here. In effect, whenever a counselor and counselee work out a plan of action together and agree to move to the execution of the plan there is an implied assumption that it will work out well—in a sense a prediction that the best possible plan has been chosen.

performance on some criterion if the test scores are to be used to predict later performances. It is important that counselors recognize the fact that predictive validity is not a necessary component of content validity or vice versa. Tests that sample current performances very effectively may fail completely to predict future performances. Failure to recognize this fact is probably one of the chief reasons why there has been so much misuse of tests. Since predictive validity is important in the use of tests in counseling, it will be given further consideration in the section on interpreting test scores in Chapters VI and VII.

CONCURRENT VALIDITY

Because it can be most readily secured, information about the concurrent validity of tests is most commonly found in test manuals. It is a relatively simple matter to give an achievement test to a group of students in any field, to collect students' marks that purport to be evidence of achievement in it, to compute the correlation between marks and test scores, and to draw the inference that the test is valid because the correlation between marks and test scores is fairly high. It is common practice to compare the scores made on a newly constructed mental ability test with those made concurrently by the subjects on a test that bears a similar label and to conclude, if the coefficient of correlation is fairly high, that the new test is a valid measure of what the other test measures.

The relative ease by which data on concurrent validity are obtained must be the chief reason for its use because other attempts to justify its acceptance by test authors seem to lack quality. If a test gives the same or very similar ranking to students as teachers' marks do, the question must be raised about the desirability of spending time and money to get a second measure of what is already available for all but a few subjects at no additional expense. It is curious to see fairly high correlation of test scores with marks

offered as evidence of validity but it is also strange to see such findings followed by suggestions that "subjective" marks now be replaced with "objective" test scores, or that a new test replace an old one simply because the new test scores correlate well with the one that was formerly used.

It appears then that counselors will not be particularly concerned with concurrent validity. A statement in a test manual that a correlation coefficient of .47 between scores on a mechanical aptitude test and current proficiency records of women operators of gum-wrapping machines is not likely to be useful to a counselor working with a boy who is trying to decide whether he should enter a mechanical occupation.³

CONSTRUCT VALIDITY

The committee of the American Psychological Association that prepared the technical recommendations for use of the psychological tests and diagnostic techniques indicated that *construct validity is ordinarily reported when the tester has no definitive criterion measure of quality that his theory implies and must use indirect measures to validate the theory. In construct validity the trait or quality underlying the test is of central importance, rather than the scores on a criterion. The concept of construct validity is vastly different from that of predictive validity in which the criterion is of utmost performance. The items on a test with predictive validity may even seem to an observer to have nothing in common with the criterion.*

Construct validity is of particular concern to those who attempt to measure personality and who, in the process, meet the problem of the lack of relationship between the personality traits they propose to measure and the overt behavior of individuals. Evidence of construct validity is commonly presented by those who attempt to

³ T. W. MacQuarrie. *MacQuarrie Test for Mechanical Ability*. Manual. Los Angeles: California Test Bureau, 1953, p. 5.

inventory counselees' interests. They depend on construct validity, when they infer that individuals have certain vocational interests when their patterns of scores on an interest inventory are similar to the scores made by members of an occupational group. Predictive validity in such cases is hard to procure because many individuals neither profess nor exhibit such interests in any situation other than the one in which they fill out the inventory.

If one is to depend on construct validity in devising a test, it is essential that the theory underlying the test be clearly defined, that the tester show how he proposes to interpret the testee's behavior, demonstrate how adequately he believes that his interpretation is justified, and demonstrate clearly the evidence and reasoning that have led him to that belief. Thus it may be theorized that persons in certain occupations have developed certain patterns of interest that are common to their group. Distinctions in the inventory patterns between such persons and all other groups would then have to be demonstrated beyond any reasonable doubt. Such theorizing and demonstration, it must be noted, could be well and thoroughly done but their completion would not necessarily mean that the patterns revealed by the interest inventory would be useful in the counseling of a youth who was in the process of choosing an occupation or training for it. To make the inventory scores useful to the counselor it would be necessary to provide data concerning the development of interest patterns by members of an occupational group, data on the consistency of their interests *before* entering the occupation or training for it, their interest patterns *during* training and early occupational experiences, and finally for a long period of work in the occupation.

It appears then that, for the counselor, evidence of the construct validity of a test will not be enough. He will look for the theoretical constructs behind any test, hope that they are sound, and look carefully at the evidence. He will then demand that some evidence of predictive validity be presented.

VALIDITY FOR COUNSELING

GROUP VS. INDIVIDUAL VALIDITY

There will always be, for counselors, the special problem raised by the fact that a test may be highly valid for groups in terms of the four categories described above and yet be invalid for the particular individual he is working with at the time. Correlation coefficients present the overall picture and the expectancy tables described in Chapters VI and VII give odds on the average, in general, on the whole, and other things being equal. The counselee may be glad to see the general relationships that they indicate, but then he may ask what the data mean *for him*. No tests provide the answer to his question. The best they can offer are generalities, odds, chances, and probabilities.

The coefficients usually presented as evidence of validity hide the fact that there may be many cases within a distribution in which *the relationships between test scores and criteria may run counter to the general trend.* Within certain subgroups of a large population the relationships may be much higher than the coefficients suggest. The general conditions under which the test was given may have influenced enough of the scores of enough of the subjects to indicate a general trend. They may not, however, have influenced the score of a particular counselee.

The counselor who is aware of the difficulties involved in the application of general findings to particular cases will realize that a coefficient of validity less than 1.00 (and most of them are far below that figure) presents *difficulties in use for counseling the individual except in terms of the odds, probabilities, and chances described in Chapters VI and VII.*

TEST INTERPRETATION FROM IMPLIED VALIDITY

Probably nothing has caused more mischief in the field of test-

ing than the naïve interpretation of validity data. Listed below are some samples from among many of the misuse of tests of questionable validity submitted by schools as evidence of the way they use tests and inventories effectively in their guidance programs.⁴

This test [The Otis Test] shows that in _____ high school competition you should earn an average grade of ____ or higher; never a lower grade than the one indicated. If you do earn a lower grade than the one recorded above you are underachieving and someone with less ability worked and traded grades with you.

Another school gives this report to a student: You should prepare for an occupation that follows one of the interests [areas from the Kuder Preference Record are given], preferably the first choice. You should plan to train in [specific occupation is cited] _____ as evidenced by results of other tests in our files.

If these were only isolated cases the problem would be bad enough but examination of reports on the use of tests in counseling suggests that they are all too common. And it seems that this kind of interpretation is due in no small measure to the overenthusiastic, exaggerated, and insufficiently qualified statements in test manuals. Consider, for example, the following quotations from such sources. (*Italics ours.*)

Word-Fluency *is* the ability to write and talk easily. People to whom words come rapidly and fluently *are* high in W. Careers requiring W include actor, stewardess, reporter, comedian, salesman, writer and publicity man. Being high in W *helps* in drama classes, public speaking, radio acting, debate, speech, and journalism.⁵

Individuals who score high in this test *possess the capacity* to under-

⁴ Subcommittee on Guidance Problems. "Extended or Potential Optimum Guidance Practices in Small, Medium, and Large North Central High Schools." *The North Central Quarterly*, October, 1949, 25:174-246.

⁵ *Self-Interpreting Profile for the SRA Primary Mental Abilities—Intermediate—For Ages 11-17*. Chicago: Science Research Associates, Revised, 1949, p. 2.

stand and profit from their experience. They *should do well* in reading literature and drama. They possess some of the basic abilities involved in understanding others and making others understand them.⁶

Studies of sales personnel, for example, indicate that a successful salesman is above average in memory, arithmetic ability, and ability to express himself well. Therefore an individual's scores on FACT tests 3, 9, and 14 (measuring his ability on these three skills) *provide an estimate of his probable success in sales work.*⁷

Those persons with high scores (on the Stenographic Aptitude Test) *can be advised* that they may enter the course with considerable likelihood of success. Those with low scores *should be counseled against this choice.*⁸

Students with profiles of this kind [better than 75th percentile standing on four factors] *possess enough of these abilities to succeed in virtually any type of academic endeavor or career* provided that interest and motivation are likewise sufficiently high.⁹

High scores on the non-verbal series should indicate likelihood of success in jobs calling for visualizing and for thinking in concrete terms. High scores on the verbal series *will indicate* probable success in jobs in which language and ideas expressed in words play a large part.¹⁰

Another approach is to note the way in which it is implied in test manuals that students' and parents' questions may be answered by use of tests. The predictive validity that is assumed in the statements is frequently unencumbered by evidence.

⁶ *Manual for the California Short-Form Test of Mental Maturity—Advanced, 1951-S Form Grades 9-Adult.* Los Angeles: California Test Bureau, 1951, p. 9.

⁷ *Counselor's Booklet: Flanagan Aptitude Classification Tests.* Chicago: Science Research Associates, 1953, p. 4.

⁸ *Manual for Stenographic Aptitude Test.* New York: The Psychological Corporation, 1939, p. 1.

⁹ *Manual for the Holzinger-Crowder Uni-Factor Tests.* Yonkers, N.Y.: World Book Co., 1955, p. 17.

¹⁰ *General Manual for the Lorge-Thorndike Intelligence Tests.* Boston: Houghton Mifflin Co., 1954.

QUESTIONS THAT
STUDENTS AND
OTHERS ASK:

STATEMENTS FROM TEST MANUALS
THAT SUGGEST THE TEST SCORE
MAY PROVIDE ANSWERS TO THE
QUESTIONS. (ITALICS ADDED)

"Should I take algebra next year?"

. . . it is probable that students ranking below the twenty-fifth percentile will find the subject so difficult that they should be excluded or diverted into a course in simplified mathematics.¹¹

"Could I be a good mechanic?"

. . . measures the ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations. This type of aptitude is important for a wide variety of jobs and for engineering and many trade school courses. The person who scores high on this trait [mechanical comprehension] *tends to learn readily* the principles of operation and repair of complex devices.¹²

"Is my son good enough in mathematics to go on in that field?"

The use of this test reveals to the student and to the teacher not only proficiency or lack of competence in the use of groups of skills, e.g., skills in the use of fractions, but also in the use of specific skills, e.g., dividing a whole number by a fraction. The test results, therefore, provide a basis for the determination of group and individual instructional needs, and the *selection and counseling of students relative to enrollment in courses and activities requiring basic skills in arithmetic.*¹³

¹¹ *Manual for the Iowa Algebra Aptitude Test*. Iowa City: Bureau of Educational Research and Service, State University of Iowa, 1942, p. 18.

¹² *Bennett Test of Mechanical Comprehension*. New York: The Psychological Corporation, 1950, p. 2.

¹³ *Manual for the Test on Basic Skills in Arithmetic*. Chicago: Science Research Associates, 1945, p. 1.

"Could my son succeed in an art career?"

Individuals falling into this range [percentiles 76-100] should, other things being equal, find *almost certain success* in an art career. Anyone making a score in the 1-25 percentiles quarter should take other available tests and inquire into all other factors mentioned above before proceeding further. If other data corroborate this finding, the person would do well to reconsider before going further into art. The individual will most likely find his abilities better suited for other lines of work.¹⁴

"Could I succeed in mechanical training?"

It is a short, single instrument capable of identifying individuals who are good risks for training in certain types of mechanical activity, and it is also an instrument which tests the ability of an individual to apply mechanical principles.

Thus a high school boy who compares very favorably with other high school boys on this test *should be encouraged* to consider seriously the mechanical field. Should he get a comparatively low score when compared with unselected high school boys, he should be encouraged to explore other occupational outlets. He *will not* be able to compete with individuals who have high mechanical aptitude if he cannot compete satisfactorily with an unselected population.¹⁵

"Have I any aptitude for clerical work?"

The paragraphs below will tell you briefly about your scores. . . . A high score on this test (*office vocabulary*) shows that you are

¹⁴ *Manual for the Meier Art Test*. Iowa City: Bureau of Educational Research and Service, State University of Iowa, 1942, pp. 9-12.

¹⁵ *Manual for the Survey of Mechanical Insight*. Los Angeles: California Test Bureau, 1955, pp. 2-5.

QUESTIONS THAT
STUDENTS AND
OTHERS ASK:

STATEMENTS FROM TEST MANUALS
THAT SUGGEST THE TEST SCORE
MAY PROVIDE ANSWERS TO THE
QUESTIONS. (ITALICS ADDED)

well equipped to learn jobs that involve letter writing, following and giving directions, reading, or talking with people. A good vocabulary is especially important for typists, receptionists, stenographers, and secretaries. A high score on this test [office arithmetic] marks the person who can learn computational jobs easily and do them well.

A high score on this [office checking] test suggests that you can master detailed clerical work easily.¹⁶

"Could I succeed in algebra?"

Every achievement test result obtained for a student during his school career has significance not only as a measure of what he has accomplished in a given course, but also *as a predictor* of what he is likely to do in the future, particularly in closely related fields.

. . . It is possible, in a given area such as mathematics, to determine whether a student is consistently strong in the field, or whether he tends to manifest greater or lesser proficiency in this area as he progresses through school.¹⁷

"What are my aptitudes?"

It is suggested that these Aptitude Tests for Occupations be given in conjunction with a standardized interest inventory. [One published by the same company is recommended.] In this way a counselor will have

¹⁶ *Profile Sheet for the S.R.A. Clerical Test*, Chicago: Science Research Associates, 1947.

¹⁷ *Manual for the Blyth Second-Year Algebra Test*, Evaluation and Adjustment Series, Yonkers, N.Y.: World Book Co., 1953, p. 6.

not only an inventory or picture of the individual's occupational aptitudes, but also a reliable source of interests. High percentile ranks *indicate the presence of aptitude*; low percentile ranks, the lack of aptitude. . . . The major difference between an examinee whose highest percentile rank is at the 85th percentile and another whose highest rank is at the 55th percentile lies in the fact that the first individual can probably handle a higher level of work in that field. For example, if Smith ranks at the 85th percentile and Jones at the 55th percentile in the General Sales field, it may be assumed Smith *will be able* to do large-scale selling (assuming equality of personality and other factors) while Jones will probably be confined to sales work behind a counter.¹⁸

"In what area in school or college can I do my best work?"

Similarly, when the tests are used for their principal purposes, the counselor can apply the results in his work with students to: (a) help the student to understand his own strengths and weaknesses in comparison with students in certain normal groups; (b) *guide the student toward choices of educational goals* and courses most appropriate for him; (c) estimate the levels of achievement to be expected of the student; (d) compare the measured academic abilities of students in different classes, grade, and school groups.¹⁹

¹⁸ *Manual for the Aptitude Tests for Occupations*. Los Angeles: California Test Bureau, 1951, p. 9.

¹⁹ *Manual for the School and College Ability Tests*. Princeton, N J: Educational Testing Service, 1955, p. 3.

PROFESSIONAL ACCEPTANCE AND ENCOURAGEMENT OF USE OF TESTS

The exaggerations and enthusiastic claims of those who would sell tests are sometimes reinforced by authors whose statements unwittingly encourage the uncritical users of tests to interpret general findings about validity as if the generalizations applied to every individual who ever took them. Samples of such statements follow: (*Italics ours.*)

A child's score on an intelligence test may be translated into a mental age which is *an index of his readiness to undertake learning tasks of a certain level of difficulty.*²⁰

In the interest inventories, the scores in many of the areas or fields have not yet been checked against adequate criteria, mainly because such criteria are difficult to establish. Even so, the *studies made in certain interest areas indicate that the scores therein are sufficiently valid for guidance purposes.*²¹

Measurements are useful for supplying *facts on which better guidance may be based.* At the present time, for example, there are *well-constructed* inventories of emotional balance, attitude scales, tests of art and music, interest blanks and procedures for measuring lying and cheating and stealing.²²

Notwithstanding these limitations the formalized measuring instruments *probably contribute more to our understanding of pupils than any other single method.* Because of the nature of their construction according to a rather rigid experimental process, they possess higher *validity* and reliability than other techniques.²³

A great deal of progress has been made during the half century since

²⁰ T. L. Torgerson and Georgia S. Adams, *Measurement and Evaluation*, New York: Dryden Press, 1954, p. 96.

²¹ J. A. Humphreys and A. E. Traxler, *Guidance Services*, Chicago: Science Research Associates, 1954, p. 136.

²² A. M. Jordan, *Measurement in Education*, New York: McGraw-Hill Book Co., 1953, pp. 4, 12.

²³ R. D. Willey and D. C. Andrew, *Modern Methods and Techniques in Guidance*, New York: Harper & Bros., 1955, p. 123.

the first serious attempts were made to measure intelligence objectively. Not only have the reliability [consistency] and validity [faithfulness of claim as to what is being tested] of the tests been increased, but a much greater variety of tests, for use under widely different conditions, now is available to the psychologist.²⁴

It is refreshing, after reading such statements as those quoted above, to find Jones saying:

A personnel officer can, by the use of tests, select the group in which most of the good material will be found; but he cannot predict among those individuals who will be successful or those who will fail. Individual prediction still eludes us, and it probably always will. This uncertainty should always be kept in mind by the counselor. The argument so often heard: "These tests are the best instruments for predictions we have and therefore we must use them" is invalid; it is based on the assumption that the counselor must know just what the client's ability is and just what he needs if adequate help is given. Such knowledge is impossible; and even if it were possible its use would violate the fundamental basis of true guidance.²⁵

Counselors, it has been pointed out, meet two problems in their use of tests. There is the common problem of securing tests that are generally valid enough so that odds or probabilities can be stated for groups. There is the additional problem of trying to interpret the scores for the person who, having accepted the general evidence of validity, says, "But what about me?" Counselor's problems with validity differ from those of the tester who seems not to regret too much that he will have some "false positives." The fact that a young man may be classified as a neurotic on the basis of a test score when there is no other evidence of any kind that he is neurotic seems to be passed over nonchalantly by those who think in terms of masses rather than persons.²⁶ That another

²⁴ L. B. Thorpe and W. W. Cruze, *Developmental Psychology*, New York: The Ronald Press Co., 1956, p. 306.

²⁵ Arthur J. Jones, *Principles of Guidance*, New York: McGraw-Hill Book Co., 1951, p. 188.

²⁶ L. J. Cronbach, *A Consideration of Information Theory and Utility Theory as*

person be rejected for training toward which he has worked for many years, and for which he may be eminently qualified except for achieving a passing score on a particular test on a particular day, does not seem to distress or disturb the authors of tests. They seem to be so satisfied with a test that selects potential achievers from a group only slightly better than does chance, and at the same time misses almost as often as it hits, that they are willing to publish and promote the test.²⁷ Their statements that tests with low validity coefficients are "far from useless" omit the statement that "use of such tests may do much harm to many individuals," even though it is equally applicable. It is always interesting to observe the behavior of persons who are willing to utilize techniques with low validity for use with other people's children but who become greatly concerned when they are used on their own progeny.

The counselor will, of course, appreciate that there are times when tests with low validity must be used because there is a job to be done hastily with large groups and with no other instruments available. During World War II, for example, it was necessary to try to select the relatively few men who were most likely to succeed in pilot training from among the many who aspired to such training.²⁸ In order to do so a battery of tests was devised, administered to the aspirants, and the predictive validity computed by comparing the test scores with their later performances in training or on the job. The predictive validity of the test battery was low but, in view of the situation and the large numbers of aspirants, the use of the test battery seemed to be justified. Counselors will not conclude because the use of a battery of tests with low validity was justified in times of war and with large numbers in a military setting that it is justified in their regular counseling duties.

Tools for Psychometric Problems. Champaign, Illinois: Bureau of Research and Service, University of Illinois, November, 1953, 65 pp.

²⁷ Psychological Corporation. "Better Than Chance." *Test Service Bulletin*, No. 45. New York: The Psychological Corporation, May, 1953, p. 5.

²⁸ J. C. Flanagan. "The Selection and Classification Program for Aviation Cadets." *Journal of Consulting Psychology*, September-October, 1942, pp. 229-239.

The authors are well aware that many of the techniques commonly used in counseling may be less valid than the tests that are available. At times it seems that it is the questionable validity of these techniques that causes counselors to seek tests because they seem to be more valid. Such faith seems not to be justified on the basis of the evidence of predictive validity contained in test manuals.

RELIABILITY

Among the many questions that arise in the interpretation of test scores there is a basic one concerning the consistency of a subject's performance. Since the items to which a student responds on a test are just a selected sample of all the questions that could be asked, it is necessary to determine how consistent he would probably be if he were asked to respond to a similar group of test items that sampled the same area of measurement. A counselor must attempt, then, to determine the extent to which a student's test score represents a true indication of his performance on the factor measured by the test. If the score does not represent a true sample and a subject's performance varies widely from one testing to another in the same area, the score will not be useful in counseling.

The common practice in testing is to present some evidence of the consistency of performances of students at one period of testing or in two testing periods separated by short periods of time. Much of the work of the counselor is, however, concerned with the problem of consistency of performances of his counselees over long intervals. He will want to know, for example, whether the student who achieved a high score on a mathematics test this year is likely to do so some years later. He will often ask, "How certain can I be that the performance on the test at the time of testing will be consistent with another performance at a much later time?" The answers to such questions will be of considerable importance in the use of test results in counseling. If the scores obtained by the

counselor are not stable he will find them of little help in evaluating current achievement levels or predicting future performances.

METHODS OF DETERMINING RELIABILITY

THE SPLIT-HALF METHOD. The split-half method is most typically used for determining the reliability of tests. To the producer of tests, this approach has several real advantages since it requires only a single test administration. It saves the time, trouble, and expense connected with methods that require a second administration of the test or the designing and administration of an alternate form. There are values and limitations to each approach. As a consumer of tests, and as one responsible for their selection and use with individuals, the counselor's concern will be primarily with the evidence of consistency presented rather than with the practical consideration of economy that may have influenced test authors and publishers.

The coefficient of reliability obtained through the split-half method is determined essentially by administering a test to a group of individuals, dividing the test into halves (usually by using the odd items as one half and the even items as the other) and obtaining scores for each half. The coefficient of correlation between the scores on the halves is computed and the result is called the coefficient of reliability. Since the reliability coefficient obtained is for a test only half as long as the test that will actually be used, the Spearman-Brown formula is employed to estimate what the reliability of the full test would be. The formula is presented in most texts in general measurement and statistics and need not be repeated here.

The split-half method is perhaps most commonly employed because of its economy. Not infrequently the reliability coefficients obtained by this method and presented in test manuals are impressively high. The counselor should examine such coefficients in de-

tail before he accepts the evidence of consistency that the split-half method purports to provide. He will need to be aware of the conditions that must be met in a test before it is appropriate to use the method and he will want to know what factors may have determined the size of the coefficients. He will need to be sure, too, that the group of individuals from whom the reliability data were obtained is similar to the group on which he proposes to use the test.

Where split-half reliability is presented in a test manual, the counselor will look for evidence that the items on each half-test were comparable with respect to content, the form in which they were presented, the difficulty of the separate items, and their range of difficulty. These conditions are assumed in the split-half method and the coefficient will be spurious and inappropriate to the degree to which they are not satisfied.

Before accepting and using reliability coefficients obtained by the split-half method the counselor will need to keep in mind, for example, that some conditions and factors that operate to cause variation in an individual's performance in the real day-to-day situation may not be reflected in the split-half method.²⁹ A person may perform variously at a given task at different times for such reasons as changes in health, fatigue, motivation, emotional strain, attention, or accuracy. Since the split-half reliability coefficient is computed from the scores on two halves of a single test, these factors will be relatively and unrealistically constant in each half. This consideration is of particular significance when the counselor is interested in prediction over a period of time that may entail normal fluctuations in performance.

The counselor must also be aware that the split-half method should not be used for speed tests. Most tests do not have items that are equal in form, content, and difficulty throughout. Such tests as the typical clerical aptitude test do, however, fall into this cate-

²⁹ These and other sources of variance of performance on a particular test are treated in detail by R. L. Thorndike in *Personnel Selection*. New York. John Wiley & Sons, 1949, pp. 72-78.

gory. Because these tests require only simple comparisons of names, series of numbers, or letter combinations, very few errors are likely to be made by individuals who take the test. Score differentiation is based primarily on the number of items completed correctly. When a test of this kind is split into two parts and scored repeatedly it is reasonable to expect the half scores to be so similar that almost a perfect positive correlation coefficient is produced.³⁰ Differences in scores of individuals are due largely to speed of response at the time of testing since the day-to-day fluctuations that account for some differentiation are not given an opportunity to operate. The more appropriate methods of determining reliability in such instances are the testing and retesting at a later time or the using of alternate forms of the same test.

Perhaps the most dramatic demonstration of the spuriously high coefficients that may result when the split-half method is applied is the now frequently cited example presented in the manual of the Differential Aptitude Test³¹ and reproduced below.

TABLE 6. Reliability Coefficients by Grade Obtained for the Clerical Speed and Accuracy Test by Testing the Alternate Forms and Split-Half Methods

Grade	Form A vs. Form B	Split-Half		N
		Form A	Form B	
8	.77	.990	.996	48
9	.83	.991	.989	50
10	.93	.996	.985	45
11	.86	.992	.993	50
12	.92	.996	.969	43

The differences and their significance for counseling an individual student will be apparent to the counselor, especially, in the

³⁰ This observation is reflected in the decision of the authors of the Differential Aptitude Test Clerical Aptitude subtest to ignore wrong answers in scoring that particular test. To quote from the manual: "In the group of two hundred and forty-five eleventh grade students, scoring for rights only and scoring by the rights minus one fourth wrongs formula resulted in only four scores which differed at all, and each of these scores differed by only one score point." *Manual for the Differential Aptitude Tests*, 1952, p. 5.

³¹ *Manual for the Differential Aptitude Test*. New York: The Psychological Corporation, 1952, p. 67.

example, at the eighth grade level. In the alternate form method, the resulting coefficient would be very questionable for group application and wholly unacceptable for individual use. If only the split-half coefficient had been reported it would have presented the test in a very favorable light. If a counselor was not aware of the limitations of the method, he might place wholly unwarranted confidence in the scores. All things considered, the counselor will do well, where he has a choice, to select those tests that employ a method other than the split-half in determining reliability. The fact that the split-half method is most economical for the publisher will not be of great consequence to the counselor whose main concern is with accuracy of measurement of individuals.

THE TEST-RETEST METHOD. The test-retest method is used frequently when there is only one form of a test. The same instrument is administered to the same group on two occasions and a coefficient of correlation of the scores on both administrations is computed. The requirement that the two tests be equivalent is satisfied in the test-retest method because the same instrument is used in each case. This advantage does not, however, guarantee accuracy of measurement. The reader will recall, in this regard, one of the questions implied in the first part of this section. It may be stated as follows, "Since the items to which the student reacted on a test were just a selected sample of *all* the questions in the same area that could be asked, how consistent would he be if he were asked to react to another group of items covering the same area of measurement?"

This is an important question, for in counseling we are very much interested in obtaining a good estimate of a student's performance in a particular area. Since in the test-retest method the sample of the "universe" is being used twice, it is impossible to cover more than one sample of a total performance. As a result an inaccurate picture may appear, for in actual work or training situations the subject will usually be required to perform in areas be-

yond those covered in the test sample. Because sampling of items is one of the sources of variance in performance, it should be allowed to operate in the test situation. It cannot do so when the same test is repeated.

The factor of memory in the test-retest method may tend to raise estimates of reliabilities and boredom in repetition may lower them. These factors operate more strongly when the lapse of time between test administrations is brief.

EQUIVALENT FORMS METHOD. A third method of determining reliability coefficients requires the administration of alternate forms of the same test. The method requires that the two forms be equivalent in type, number, and level of difficulty of the items. If it is reported in a test manual that two forms of the test were constructed at the same time from the common pool of items that was used in a preliminary tryout, the counselor can be relatively certain that these requirements are met. In general, the alternate-forms method will provide for the counselor the most usable estimate of reliability of the three major methods. Since the method requires that the tests be given with a time lapse, the day-to-day fluctuations of behavior of the individual will have a chance to influence test performance as would be the case in actual work or study situations. Further, because the two tests represent two samples of the total area measured, they offer a more complete check on performance. The counselor has at least a partial answer to the question about whether the student's performance is representative of what he really can do with the types of problems and questions presented in the test.

THE KUDER-RICHARDSON METHOD. A final method applied in the determination of test reliability utilizes analysis of variance procedures to determine the consistency with which subjects respond to the test items. While several formulas based on this principle have been derived, the one most widely applied is known as the Kuder-Richardson.²² It has some of the limitations of the split-half

²² For a detailed discussion of this method, see Robert L. Thorndike, "Reliability,"

method in that it cannot be applied to speed tests. Moreover, it does not reflect the day-to-day variance that might be found in an individual's performance, because it was computed, as in the split-half method, from a single test administration. Like the split-half method, then, it fails to afford the desired evidence of consistency over a period of time.

FACTORS INFLUENCING RELIABILITY

In the brief discussion above, an attempt has been made to describe the methods, limitations, and advantages of the methods for determining the reliability of tests most commonly employed. Some additional conditions to which the methods described are sensitive and that will be important in the interpretation and use of reliability data are presented below.

Reliability coefficients are affected by the range of talent represented in the group to which the test is administered. In general, the wider the range of talent represented within the group whose scores are used in determining reliability coefficients the higher they will be. If the counselor accepts the coefficient reported in the manual he assumes that the group on which he is going to use a test can perform over similar ranges as the group on which the reported coefficient was obtained. This means, as a minimum, that the manual should indicate the number, age, sex, and educational levels of the subjects from whose scores the reliability coefficients were computed.³³ While it is seldom done, it would further enhance the accuracy of the reliability data if several independently obtained reports of test reliability from various geographic areas were presented.

The effect that range of talent can have on reliability data is suggested by Travers:

in E. F. Lindquist (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 586-594.

³³ Technical Recommendations for Psychological Tests and Diagnostic Techniques, *Psychological Bulletin Supplement*, March, 1954, p. 32.

A test developed to measure knowledge of the names of tools consisted of 120 multiple choice test items. When this test was administered to all the high school seniors in a small city, it was found to have a split-half reliability of .95 and the scores ranged from 20 up to 115. However, when the same test was administered to a group of machinists, the scores ranged from 106 to 117 and the reliability estimated from this group was .20. In the case of the machinists, differences in scores would be considered almost meaningless, and this is reflected in the low reliability coefficient derived from this group.³⁴

It can be seen from the example that the counselor cannot afford to be enticed into playing guessing games with a test manual and that he cannot assume that a reported coefficient can be applied with confidence to any group of subjects.

The reliability of a test may not be equal at all parts of the range that the test is designed to measure. If the test is designed to measure a particular skill or performance over Grades 9 through 12, the reliability coefficient may vary at each grade level. An example of differences in the reliability estimates found among grades on a new test of clerical aptitude³⁵ is presented in Table 7.

The reader will note that the reliability coefficients on the verbal skills subtest range from .73 for ninth graders to .90 for a combination group of eleventh and twelfth graders. While the coefficient of .90 for eleventh and twelfth grade students is fairly adequate for individual counseling, the .75 for the ninth graders might prove to be unsatisfactory for that purpose. It may be noted that the coefficients reported for the two groups of ninth graders and two groups of eleventh graders are different, and, in the case of written directions for ninth graders, the difference is .10. The

³⁴ Robert M. W. Travers, *Educational Measurement*, New York: The Macmillan Co., 1955, p. 71.

³⁵ *Manual for the Turse Clerical Aptitudes Test*, Yonkers, N.Y.: World Book Co., 1955, p. 10.

TABLE 7. Split-Half Reliability Coefficients

Measures	Grades	N	r
Verbal skills	9	50	.73
	9	53	.77
	11	47	.88
	11	48	.80
	11-12	320	.90
Number skills	9	50	.82
	9	53	.87
	11	47	.94
	11	48	.88
	11-12	215	.83
Written directions	9	50	.77
	9	53	.87
	11	47	.89
	11	48	.88
	11-12	236	.85
Learning ability	9	50	.88
	9	53	.90
	11	47	.93
	11	48	.89

tests seem to be most reliable at the eleventh grade level. The manual indicates that, since the coefficients were obtained by the split-half method, they may be slightly overestimated.

The number of items in a test may affect the size of reliability coefficients. In general (though there is a point beyond which this may not be true) the longer the test the more reliable it will be. This factor will not ordinarily be of much concern to the counselor when a test yields a single score. It will be of importance, however, when a test is designed to provide subtest scores as components of the overall function of the test. This is the case in some "diagnostic" tests. Of importance to the counselor is the fact that the reliabilities of these subtests are nearly always low and frequently too low to be used for differential diagnosis. This is true largely because the relatively few items in the subtests provide an inadequate sample of the total number that might have been used. It is important, then, that subtest reliabilities be reported in addition to that of the test as a whole. If only the latter is reported the

counselor must not attempt to interpret or use the subtest scores. In this regard, it is refreshing and encouraging to find the following note in the manual of a test just recently introduced. "The user of the School and College Ability Tests cannot be reminded too emphatically that the scores of individuals on the four subtests should *NOT* be interpreted separately. The part scores and total scores for which the interpretive materials provide recording spaces and normative data *are* reliable enough for individual use, but separate subtest scores should *never* be recorded for individuals." ³⁸

The appearance of such a warning in a test manual is an encouraging sign that we may look for increasing responsibility on the part of test publishers. Unfortunately, the example is the exception rather than the rule. It is not uncommon to find test manuals encouraging interpretation of reliability estimates far beyond those permitted by the data.

Such factors in the administration of tests as instructions, timing, and motivation may influence the reliability of tests. When a person accepts reliability coefficients he assumes that the data were obtained by uniform administration of the test. He must also assume that the reported estimate will not be obtained if he departs from the instructions in the manual.

It may be obvious to the reader by this time that, with many factors operating to influence scores and consequently reliability estimates, a perfect reliability coefficient of 1.00 is never obtained in educational and psychological testing. It may also be seen that, depending upon the degree of error in measurement, the scores obtained on students vary so much that "true" scores are never obtained.

A "true" score could theoretically be obtained if we gave a student an infinite number of samples of the task to perform and averaged his scores. Because coefficients are not based on true scores, but rather on those that are obtained at the time of sam-

³⁸ *Manual for the Coöperative School and College Ability Tests*. Princeton, N.J.: Coöperative Test Division, Educational Testing Service, 1955, p. 11.

pling, reliability coefficients may be depressed. This fact has occasionally prompted test designers and publishers to correct the reliability for "shrinkage due to errors" and the counselor may, therefore, encounter reliability coefficients that have been "corrected for attenuation." This correction means that the reliability has been computed on the basis of estimated true scores rather than obtained scores. Thus, the errors of measurement that are present in actual scores are reduced or theoretically eliminated and the resulting coefficients are likely to be considerably higher than those obtained when actual scores are used. Thorndike summarized the implications of the above very well: "Practical prediction must be done with existing fallible tests. To some extent it is misleading to present corrected correlations between hypothetical true measures. The prediction which could be achieved by a hypothetical, perfectly reliable test may be quite misleading because such a test is never available to us." ³⁷

Among other things, since day-to-day fluctuations that may affect test performance are present in human performances there seems little justification to remove them from the test situation. If corrected coefficients are presented, the uncorrected coefficient should also be indicated.

RELIABILITY OF INDIVIDUAL'S SCORES

The reader acquainted with statistical procedures may have noted by this time that the coefficients of correlation produced by use of the methods described above provide a limited amount of data regarding a particular student's score. The coefficient, for all practical purposes, indicates the degree to which individuals comprising a group maintain their positions from one test administration to another but it does not tell us about the consistency of individual's scores. Since counselors make predictions and take

³⁷ Robert L. Thorndike, "Reliability," E. F. Lindquist (Ed.), *Educational Measurement*, Washington, D.C.: American Council on Education, 1951, p. 613.

action on the basis of the scores earned by one student at a time, they will be more concerned with the accuracy of measurement of an individual than of groups. If a student obtains a score of 75 on a given test (assuming that the performance represented by the test score has implications for some future performance and the implications vary with the magnitude of the score), the counselor will want to know how consistently the student will score at that level. He will know that the subject is not likely to be completely consistent and his question becomes "How much will it vary?" If the scores achieved by the same individual on repeated measurements varied widely, he could not know what significance to attach to it in terms of future performance on a job or in training at a later time. This being the case, the counselor should demand some evidence about the relative stability of the score with which he is working.

He will find some of the evidence he seeks in the standard error of measurement. It is an estimate of the amount by which an obtained score is likely to vary from the individual's true score. The SE_M , usually presented in terms of raw scores, indicates the range between which persons' scores are likely to fall on retesting. Thus, if the reported SE_M for a given test was five raw score points and a student obtained a score of 75, one could not be sure what *he* would do on a second try at the test. The SE_M would *not* indicate what a particular person would do on a retest.

A table, reproduced in part from the manual ²⁸ of a new Clerical Aptitudes Test, may suffice to illustrate the use of the standard error of measurement.

The counselor will note in Table 8 that the median standard error of measurement obtained from those computed for six different groups (presumably one group for each of Grades 9, 11, and 11-12 combined, tested, and retested for each measure) on verbal skills was 2.7. The reader will note further, however, that the 2.7

²⁸ *Manual for the Turre Clerical Aptitudes Test*. Yonkers, N.Y.: World Book Co., 1955, p. 10.

TABLE 8. Range and Median Standard Errors of Measurement

Measure	Number of Groups	Range *	Median	In Percentile Terms
Verbal skills	6	1.8- 3.1	2.7	16
Number skills	6	1.8- 2.3	2.2	15
Written directions	6	1.7- 2.4	2.0	21
Learning ability	6	3.6- 5.0	4.6	13
Clerical speed	6	5.4- 7.7	5.8	16
General clerical Aptitude	6	8.2-14.0	11.7	12

SE Meas $\approx \sigma \sqrt{1-r}$ based on groups on which test-retest and split-half reliability coefficients were computed.

* In raw score terms.

figure is the median on the verbal skills subtest obtained from ninth graders as well as tenth, eleventh, and twelfth graders combined. While the range of SE_M represented (1.8 to 3.1) may not mean a great deal in interpreting the scores in this case, it is obvious that the standard errors of measurement vary with the different groups and that the *median* SE_M of 2.7 may be somewhat misleading if applied equally to all groups. The significance of the SE_M of the example can be seen in the figure "16" under the heading "In percentile terms." When it is taken at or near the mean, the error (2.7) of the obtained score may be converted to percentile points on the table of norms provided. From the data we can be reasonably sure that the score will place persons on the average between plus and minus 16 percentile points of the percentile indicated by the obtained score.

It can be seen that, for a given grade and SE_M , the range of percentiles in which scores are likely to fall will vary. It must also be noted that the magnitude of the SE_M in terms of raw score points may not tell the whole story even for a group, for, referring again to Table 8, the lowest SE_M , 2.0, on written directions, has the highest percentile range, and it is from the percentile rank that we ordinarily draw the implications. The results of such computa-

tions fail, however, to indicate how consistent a particular counselee will be.

LONG-TERM STABILITY

Before closing this discussion of reliability it may be of value to consider the additional problem of stability of scores over long periods of time. Much of counseling is concerned with the problem of variability in later performances of counselees and, when counselors use tests, they do so in the hope that the results will help with this problem.

The reader will recall from the first chapter that counselors are asked such questions as these: Do I have what it takes to succeed in college? What curriculum should I follow when I enter high school next year? What is the likelihood that this student will complete an apprenticeship in the machinist trade? These questions refer to performances that may range from months in the time between the end of eighth and the beginning of the ninth grade to a span of several years.

The need for some long-term evidence of stability of test scores has not been entirely ignored in the history of testing. Traxler recognized the problem in intelligence testing some twenty years ago when he commented:

The reliability of intelligence quotients derived from tests administered at least a year apart is, in all probability, lower than it would be if the tests were given with only a short time interval between them. This situation is to be expected unless the IQ is perfectly constant. However, in view of the fact that the IQ's found for a class at a time of entrance to high school are frequently recorded and used for several years by the school, the correlation between the forms of tests administered in successive years is a matter of considerable practical importance.³⁹

³⁹ Arthur E. Traxler, "Reliability, Constancy, and Validity of the Otis IQ," *Journal of Applied Psychology*, 1934, 18:243-244.

Thorndike anticipated the problem more recently when he stated that "of course, for some purposes we may be interested in consistency of performance over an extended period of time, but consistency of this type represents a rather difficult concept of reliability."⁴⁰

He writes further to the point when he suggests that "for use in connection with predictions and evaluations extending over some period of time, the meaningful procedure appears to be to retest with a similar time interval."

Referring to this concept as "stability of scores," the committee that prepared the recent *Technical Recommendations* regards as "essential" that "the manual should indicate what degree of stability of scores may be expected if a test is repeated after time has lapsed. If such evidence is not presented, the absence of information regarding stability should be noted."⁴¹

It further observes that "most educational and psychological tests measure qualities which are presumed to be stable for some time, unless training or specific experience intervene."

While it has taken a long time for a recognition of this concept to be translated into application, there is a suggestion that, with the appearance of new tests, the counselor will have some helpful data. A recent example is found in the excellent manual accompanying the School and College Ability Tests. This manual, following closely the *Technical Recommendations*, contains the following statement: "The coefficient of stability, a correlation between scores earned by the same students on different forms of the test before and after a substantial period of time often is a useful measure of the stability of scores. Although the SCAT series measures abilities that are *expected to change* under good instruc-

⁴⁰ R. Thorndike in E. F. Lindquist (Ed.), *Educational Measurement*. American Council on Education, Washington, D.C.: 1951, pp. 571-617.

⁴¹ *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, p. 32.

tion, the stability characteristics are under study. No stability data are available at the time of publication, however."⁴²

It may be encouraging to the counselor to know that some publishers are cognizant of the need for such data and that, while not commonly available at present, it may be forthcoming. In the absence of such data it is, of course, possible for counselors to do a little experimentation of their own.

The use to which a counselor puts test results will determine the need for information on long-term stability of scores. If, for example, test results are used to put students into subgroups of a class that is already formed, it is unlikely that such long-term data will be necessary. If, on the other hand, the counselor expects to use the results, as one recent test proposes, "To aid students at the eighth or ninth grade levels in the selection of appropriate high school programs,"⁴³ the long-term reliability may be of considerable importance.

There seems to be little question that the production of tests yielding relatively stable scores over an extended period of time is indeed a challenging task. It may require new approaches to measurement. It seems, however, that the task of counseling is not likely to be altered significantly in respect to the use of predictive data; but it is reasonable to expect that testing instruments can be devised to help in the problems prediction entails. While the counselor waits for the millennium his consideration of the problems may remind him that he is working with very fallible instruments.

RELIABILITY DOES NOT GUARANTEE VALIDITY

A counselor may wish for perfect instruments, but he must realize that he does not have them. The next best thing is to be aware of the limitations of those he does have. He will use his knowledge of the limitations of reliability data in selecting a test and in apply-

⁴² *Manual for the School and College Ability Tests*. Princeton, N.J.: Coöperative Test Division, Educational Testing Service, 1955, p. 12.

⁴³ *Turse Clerical Aptitudes Test*. Yonkers, N.Y.: World Book Co., 1955, p. 1.

ing the test results in counseling. In doing so, he will, of course, keep in mind that with the best reliability data currently available he still may not have what he wants, evidence that the test he has selected will actually do what it purports to do. In the usual evidence of reliability presented in test manuals the counselor merely has an indication of the consistency with which the test currently measures something. But he must look to evidence (discussed previously under *validity*) that the performance represented by the score has meaning and implications in terms of the problems with which he and his counselees are working.

It seems most unfortunate that the term "reliability" has been appropriated from common usage and employed as a technical term in measurement. To many persons reliability connotes dependability, trustworthiness, and generally high value. It should be noted that a highly reliable (in the technical sense) test may be useless in counseling. The counselor will always look beyond the elaborate tables of reliability coefficients frequently found in the catalogues of test distributors to find evidence of high predictive validity. If he finds the latter he need be less concerned with the former.

NORMS

Discussion of the various forms in which normative data appear may be found in any elementary textbook on educational measurement, and brief statements about the merits and limitations of the several forms of derived scores appear in Chapter V of this volume. It is suggested there that percentiles, despite the fact that they do not represent equal units on a scale, are probably the most useful of all derived scores in a counseling process requiring that results of a testing program be interpreted to counselees, their parents, and school personnel.

To point out some of the difficulties of test interpretation from norms given in test manuals, the case of Joe is given in some detail.

He was an eleventh grade boy who was trying to choose between a training program for engineering and one for auto mechanics. He had achieved the following raw scores on the Differential Aptitude Test battery: ⁴⁴

TABLE 9. Differential Aptitude Scores of Counselor Seeking Guidance

Subtests	Joe's Scores
Verbal Reasoning	21
Numerical Ability	20
Abstract Reasoning	33
Space Relations	72
Mechanical Reasoning	63
Clerical Speed and Accuracy	32
Language Usage: I Spelling	2
II Sentences	4

A quick glance at the absolute size of each score might lead to the drawing of some false conclusions. It would appear from the above figures that Joe's performance on the space relations test is better than on the mechanical reasoning test. The numerical ability score, which is smaller than the clerical score of 32, suggests that he performed more effectively on the latter test. These raw scores are not meaningful until his scores are compared with the 2,700 eleventh grade boys whose scores are reported in the test manual. Using these data for comparisons in terms of percentiles, the counselor would see Joe's standing as follows:

TABLE 10. Differential Aptitude Scores as Percentiles

Subtest	Joe's Scores	Percentiles
Verbal Reasoning	21	35
Numerical Ability	20	50
Abstract Reasoning	33	55
Space Relations	72	80
Mechanical Reasoning	63	97
Clerical Speed and Accuracy	32	3
Language Usage: I Spelling	2	5
II Sentences	4	3

⁴⁴ *Differential Aptitude Tests*. New York: The Psychological Corporation, 1952, pp 25-34.

The use of such percentile norms permits the counselor to compare Joe's scores with those of other high school boys of similar grade levels, and to make some assumptions about his relative strengths and weaknesses on the factors measured by the subtests of this particular test battery. Joe's performances might have been compared with other groups with known characteristics, such as engineering students, engineers, auto mechanics, or college freshmen. In each case his percentiles would differ from those obtained by comparison with high school boys at his own grade level.

CHARACTERISTICS OF NORM GROUPS

Before the counselor can use the percentiles presented above, he must become thoroughly familiar with the selection and description of the *normative groups* used by the authors of the test. He would need information about the size of the standardization population and about the subject's socioeconomic level, geographic location, and educational experience. If he is counseling with the boy about some vocational or educational decision, he must know about the performances of members of a group who have entered the kind of training that Joe is considering and about the test performances of some successful members of the occupation. He will seldom find adequate information about such matters in test manuals.

To help Joe to answer his query about whether his test performances are most like engineers or auto mechanics, the most meaningful norms would be those that permitted comparison of his scores with those of similar students who had considered such occupations while they were in high school, had subsequently completed such training, and had entered the occupations. Such norms would provide comparative information at the high school level and at several strategic points in training and employment. And those data should be based on the continuous testing of performance of *the same groups before and during training*. Since no sizable body

TABLE 11. Types of Norms Reported in Manuals of Representative Aptitude Tests

Test	Type of Norm Provided	Size of Norm Groups
Differential Aptitude Tests	1. Educational: Grades 8-12. 2. Sex.	1. Ranges from 2,100 for 12th grade boys to 7,400 for 9th grade girls.
Test of Mechanical Comprehension (Bennett)	1. Educational: Grades 9-13 for men. 2. Occupational: applicants for jobs. 3. Women: college freshmen, trainees and applicants.	1. Ranges from 300 12th graders to 833 9th graders. 2. Ranges from 143 candidates for engineering jobs to 2,217 applicants for job of mechanic's helper. 3. Ranges from 111 college freshmen to 1,090 trainees in an airplane factory.
Minnesota Paper Form Board (Revised 1948)	1. Educational: Grades 10-12 through 5th year engineering. Includes some age and sex groups. 2. Occupational: by sex, includes shop work applicants to Time and Motion Study engineers. 3. Geographic: New England 11th and 12th graders, Minneapolis 9th graders, Illinois Institute of Technology freshmen, etc.	1. Ranges from 143 male engineering juniors to 1,288 12th grade boys. 2. Ranges from 119 men in a leadburner's course to 994 male prison inmates. 3. Ranges from 178 9th grade girls to 46,943 9th grade boys and girls.
MacQuarrie Test for Mechanical Ability	1. Age: 10 years to adult. 2. Sex.	1. Size of groups not given. 2. 1,000 males and 1,000 females.
SRA Mechanical Aptitudes	1. Educational: 9th to 12th by sex. 2. Occupational: male trainees for mechanical occupations.	1. Ranges from 298 11th grade girls to 2,240 9th grade boys. 2. 650 male trainees.

of such norms exists for any test today, the counselor must consider several alternatives.

He might, for example, use the tenuous assumption that there

are no important differences in crucial psychological and educational factors between high school subjects and groups in training or on the job. In this case he would make comparisons of Joe's performances with those of such groups. Authors of several commonly used tests that are said to provide assistance in making vocational decisions offer norms for many groups whose characteristics are known. Examination of the list of norms provided for such tests (on p. 84) suggests that the authors have decided that educational, occupational, age, and sex groups are different enough to need special norms.

INTERPRETING NORM DATA IN COUNSELING

While an array of test scores obtained from all the tests given in Table 11 might be far more than a counselor in a public high school would have at his disposal, it is interesting to speculate how he could use them in counseling with Joe. The emphasis at this point is on norms and it will be assumed for the moment that the criteria of validity and reliability, as discussed above, have been as adequately met as current tests can do so.

The Differential Aptitude Test scores previously mentioned are based on the test performances of 2,700 eleventh grade boys. The counselor has been given no information about the tentative or actual vocational plans or specific educational histories of members of this group. Without them Joe can only be shown that, in comparison with this group of his peers, his performance on the mechanical reasoning subtest is equaled or exceeded by only 3 percent, and his performance on numerical ability is average. If the counselor applies standard error of measurement concepts rigorously, even less can be told of this particular counselee's performance except that, in the areas *assumed* to have meaning for potential engineering students, his scores lie in the middle and upper ranges of eleventh grade boys.

A bit more interpretation is possible in view of evidence about subsequent progress of small numbers of these 2,700 students who entered training for the field of engineering or entered engineering or allied occupational fields. Bennett⁴⁵ reported that 53 subjects who had completed engineering training by 1955 had scored on the average at the 87th percentile on the numerical ability section of the test; at the 77th percentile on space relations; and at the 82nd percentile on the mechanical reasoning test while still in high school. The average high school percentiles of the 22 men actually engaged in engineering in 1955 were 89 on numerical ability, 81 on space relations, and 86 on mechanical reasoning. A further breakdown of the scores of these 22 employed engineers indicated that their average range of scores while in high school in 1947 encompassed numerical ability scores between the 60th and 95th percentiles and their space relations scores ranged from the 55th to 95th percentile. Taking this variability into consideration, Joe's poorest showing, his 50th percentile on the numerical ability subtest, falls just short of the attainment of this successful group. His other scores, assumed to have something to do with chances for success in training for engineering, compare favorably with the 22 employed engineers.⁴⁶ Should the counselor now encourage Joe toward engineering by stating that his chances for success are probably above average and point up the similarity of his score to those of the 22 men? Or should the counselor, recognizing the rather small numbers of students involved in this comparison and, admitting that he knows next to nothing about them except that they attended high school in 1947 and entered the engineering profession in 1955, be more cautious in his interpretations to the counselee?

Perhaps, when he recognizes the difficulties noted above, the

⁴⁵ George K. Bennett, "The DAT—A Seven-Year Follow-Up," *Test Service Bulletin* New York: The Psychological Corporation, November, 1955, p. 8.

⁴⁶ Consideration of percentiles of 5 on spelling, 3 on sentences, and 35 on verbal reasoning might raise the question of the boy's all-around chances for success in college training with its heavy emphasis on verbal skills.

counselor will attempt to get more test scores. As he begins to do so, further problems in norms arise. With the Differential Aptitude subtests he had a clear-cut standardization group whose members were similar to his subject. With some other tests he must compare the boy's performance with diverse and unlike groups selected at widely varying times and places, and under many and varied circumstances. For example, on the Minnesota Paper Form Board,⁴⁷ a well-known spatial relations test which is commonly used in prediction of success in technical, mechanical, and engineering training, the counselor has the choice of 13 educational and 13 industrial groups for comparison purposes. The size and composition of these groups vary from 119 leadburner trainees in Delaware to 1,123 ninth and tenth grade superior boys and girls, who were clients of the Cleveland Jewish Vocational Service.

If the counselor compares Joe's raw score of 45 with the first of these groups, the leadburner trainees, he finds that it lies between the 80th and 90th percentiles; a comparison with the Cleveland boys and girls places him at the 70th percentile. He may also compare the boy's performance with 334 male International Business Machines customer engineers at Endicott, New York, and find that his score places him at the 35th percentile. Perhaps the most appropriate available norms for this particular case would be the 344 first year engineering students at Northeastern University. Joe's scores place him at the 60th percentile for that group. The data were gathered prior to 1941 and it is, therefore, likely that the characteristics of first-year engineering students and engineering curricula have changed, particularly since the beginning of the Korean War and the consequent strong recruiting drive for engineering students. And since the performance measured by the Minnesota Form Board is just one variable that is frequently regarded as necessary for successful training in engineering, it is likely that the counselor and the counselee will become discour-

⁴⁷ *Manual for the Revised Minnesota Paper Form Board Test*, New York: The Psychological Corporation, 1948, pp. 12-15.

aged in their quest for a simple yes or no answer from tests to the question: "Do I have enough talent to train as an engineer?"

As the next step, the counselor might attempt to use scores on the Science Research Associates Primary Mental Abilities test to help Joe to answer his questions. Here his performance on Number, Space, and Reasoning subtests—all subtests that the manual implies would be helpful for prediction in this area—were at the 65th, 50th, and 80th percentiles respectively on the age norms given. Reference to the manual for this test reveals that the norms were based on the performance of 18,000 high school students. Unfortunately, this group was not further described or differentiated by sex, location, achievement, socioeconomic position, or vocational plans. Research⁴⁸ has shown that there are clear and important sex differences in scores on this battery that should be taken into consideration. Even with supplementary information about normative data gleaned from research literature, the counselor's interpretation of scores on this test must be limited.

Specific and useful norms relating to the questions about prediction of success in training can be answered only if norms intended to permit comparisons with students in general, students in specialized training, successful entrants to the occupation, and later performance by members of the occupational group are provided. Such data could be produced by longitudinal studies if sizable groups were followed through their preparation, training, and work experience. On the reverse side of the "self-interpreting" profile of the Primary Mental Abilities test, the counselee may read that an engineer needs the "ability" to visualize objects in space and the subtest "Space" measures this quality. This kind of oversimplification and self-interpretation is highly misleading and dangerous. The counselor who looks for evidence to justify this statement will not find any and having done so he should reject the test.

⁴⁸Frederick Herzberg. "A Study of Sex Differences of the Primary Mental Abilities Test." *Educational and Psychological Measurement*, 1954, 14:687-689.

Little will be gained from appraisal of Joe's scores on the Science Research Associates Mechanical Aptitudes Test where the only comparison of his performance that may be made is in terms of 650 male trainees, high school graduates, nongraduates, and veterans who were in training for mechanical jobs as apprentices. No other data about the norm group are given!

More detailed norms are provided for the Bennett Test of Mechanical Comprehension (Form AA).⁴⁹ Here the counselor has his choice of educational norms that present the performance of eleventh grade male high school students, technical high school seniors, and engineering school freshmen. Occupational norms are offered for candidates for engineering positions, men in the Works Progress Administration mechanical courses of depression days, trainees in an airplane factory, applicants for jobs of mechanic's helper, and several other industrial groups. Detailed and specific descriptive data about these groups and information pertaining to their selection are missing.

Joe's score of 44 may be converted into percentiles for each of these norm groups as follows.

TABLE 12. Conversion of Scores to Percentiles

<i>Group Compared</i>	<i>Percentiles</i>
Eleventh grade male high school students	75
Technical high school seniors	65
Engineering school freshmen	30
Candidates for engineering positions	15
Men in WPA mechanical courses	90
Trainees in an airplane factory	55
Applicants for jobs as mechanic's helper	75

Now Joe and his counselor may become involved in a series of mental gyrations as they attempt to choose the most meaningful group or groups with which his performance should be compared. One might be tempted to generalize to say that a comparison of

⁴⁹ *Manual for the Bennett Test of Mechanical Comprehension*. New York: The Psychological Corporation, 1950.

Joe's performance with that of other eleventh grade high school boys indicates that one quarter of them equaled or exceeded his score. He might also be tempted to say that, since comparison of his test performance with that of engineering school freshmen shows that nearly three quarters equaled or exceeded his score, there would be serious doubts of his "measuring up" in engineering training. Yet the counselor cannot be sure of this since the two groups used for this comparison were selected at different times and in other places. The counselor has little knowledge of important training, educational or motivational factors in the individuals composing these groups, and he lacks information about the selection, achievement, or mental level of the members. It may be indicated generally that, since the score of 44 approximates successively lower percentiles as the comparison is applied to groups with increasingly more training (i.e., engineering freshmen, candidates for engineering positions, etc.), the factors of training and selection become increasingly important. There are insufficient data, however, about the composition of these groups to warrant the acceptance of the generalization. Joe and his counselor must seek more data to help in making decisions.

In the manual of the MacQuarrie Tests for Mechanical Ability⁵⁰ one may find the statement that ". . . [it] has been used to measure the aptitudes of more than 5,000,000 persons. . . ." This impressive number of users must have had difficulty in interpreting their scores, since only undefined age and sex norms are provided for the instrument. Some attempt has been made in subsequent studies⁵¹ reported by the publishers to present "tentative" standards of performance for groups of operators of gum-wrapping machines, sewing-machine operators, leather workers, and aircraft workers. Careful examination of this body of norms fails, however, to indicate the numbers used, selective factors involved,

⁵⁰ *Manual for the MacQuarrie Test for Mechanical Ability*. Los Angeles: California Test Bureau, 1933.

⁵¹ June C. Duran. *MacQuarrie Tests for Mechanical Ability. Summary of Investigations*, Number 2. Los Angeles: California Test Bureau, 1930.

geographic location, or any other pertinent information about the subjects on which the norms were based.

LOCAL NORMS

When test publishers fail to provide adequate norms the counselor may establish local norms. This task can usually be accomplished over a long period of time by amassing enough cases. It is a difficult and expensive process but it may be necessary to supplement and give added meaning to the scant normative information furnished by most test publishers. Often, because of peculiarities of the population with which he is working in relation to locale, ethnic origins, educational level, mental abilities, or other selective factors, the counselor will find his groups considerably above or below so-called national norms.

Two examples illustrate this point. While one of the authors served as guidance director of a midwestern university high school he noted that the students, who were a highly selected group from superior socioeconomic class homes, scored a year or more above published norms on standardized achievement tests. The staff of this school prided themselves on this fact. Comparison of these students with others by use of Educational Records Bureau norms for private schools indicated, however, that these students were just at grade level when compared with their peers. In this case the development of local norms made realistic comparisons possible.

The teachers in an Indian Service boarding school in the southwest were greatly concerned because their charges continually tested at one or two years below published national norms on standardized achievement tests. Comparison of these same students with Indian Service norms showed that they were fully one year ahead of other Indian Service school students throughout the country. Less spectacular examples might be cited in numerous instances.

More frequent use of local norms would permit some control over the circumstances of testing, the conditions of set, expectation, motivation, reward, and other psychologically important factors that must be weighed in the interpretation of scores. Norms that would differentiate between testees who took the test under routine test administration procedures might differ from those that were developed only upon highly selected and highly motivated testees. The difficulty in making any sort of a meaningful comparison between generalized high school grade norms and those of selected groups such as engineering freshmen or candidates for engineering positions is a case in point. Members of the former group produce a wider range of scores as evidenced by larger standard deviations. It must be recognized also that such important variables as attitude, effort, attention, or motivation are largely unstandardized no matter how many "standardized" time limits and directions are used. The effect of these factors on norms developed from such highly specialized groups of candidates, selected students, or successful trainees is such that the counselor is frequently required to make comparisons of unlike subjects. This process is not likely to improve his counseling.

ADMINISTRATION OF TESTS

The manner in which tests are administered affects subjects' performance and influences the interpretation of their scores. It would indeed be helpful if the counselor could be assured that the score he had obtained represented the students' best efforts but this cannot always be taken for granted. The counselor must be aware of the factors that may have been operating to make the test performance less than best and he must attempt to control these factors when he is giving a test.

The status of administration, among factors related to the use of psychological tests, has been well described by Traxler:

There is, however, an equally important area (as constructing scoring,

and the use of objective tests) in the whole process from the building of a test to the application of the results in individual guidance that is largely neglected. This area is the administration of tests. It seems highly inconsistent for the producers and consumers of tests to expend much time and energy in order to obtain the best possible tests, to score them accurately, and to use the results professionally, and at the same time to treat very casually one step in the process, which, if incorrectly carried out, can invalidate all the rest of the work.⁵²

Writing in the same vein more recently, Traxler restated his position in these words:

In view of their crucial importance in the whole chain of events from the conception of the test to the use of the scores in conferences with individuals, it seems highly unfortunate that the giving and scoring of tests are frequently treated very casually by both the authors and the users of tests. Test specialists have been very dilatory in providing research data on the many debatable points relative to test administration and scoring, and, in general, test makers have not applied the same care and zeal to the writing of directions for administering tests that they have applied to item validation and other technical aspects of test construction.⁵³

The space and treatment given to the factor of test administration as contrasted to other aspects of use of tests in two recent parallel publications of major importance in the testing movement attest to Traxler's observations. Both the Technical Recommendations for Psychological and Diagnostic Techniques⁵⁴ and Tech-

⁵² Arthur E. Traxler, "Needed Improvement in the Administration of Objective Tests." *Fourth Yearbook of the National Council on Measurements Used in Education*. Fairmont, West Virginia: The National Council on Measurements Used in Education, 1947, p. 3.

⁵³ Arthur E. Traxler, "Administering and Scoring the Objective Test," in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, p. 329.

⁵⁴ *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, D.C.: American Psychological Association, 1954.

nical Recommendations for Achievement Tests⁵⁵ devote only a few short statements to desired standards regarding test administration. They suggest to the reader that good administration of the testing program is taken for granted, or that this particular aspect of test usage is of relatively little importance. Inspection of many test manuals must lead one to conclude that these assumptions are similarly held by test designers and publishers. It does seem, however, that the implications of test administration are such that it warrants considerable attention by test users.

In one sense, of course, the tendency to gloss over the importance of the administration of tests is understandable. Proper administration in itself does not contribute directly to the value of a test. Many things can happen if the administration is not good, but optimal administrative practices do not enhance the basic value of the test. Poor administration can have a lowering effect on the validity and reliability of test results, but optimum administration in a given situation does not improve these factors above the limits imposed by the basic data.

Details of test administration can be something of a nuisance and it is easier to avoid such questions as the following that might come to mind. Did the student fully understand what he was to do in taking the test? How much of the score was due to guessing? Could he do better if more time was allowed? How important is the factor of time for this student? How apprehensive was the student about taking the test? Did the apprehension influence the scores and if so, how much? How well motivated was the student? Does the score represent the best he could do? Was the student's physical status normal? How diligent was he in pursuing the test items? How much time was consumed by daydreaming, looking out the window, or reacting to distractions?

As Traxler has pointed out, research has not supplied answers to many of the questions that might be asked about optimal test

⁵⁵ *Technical Recommendations for Achievement Tests*. Washington, D.C.: American Educational Research Association, 1955.

administration procedures and one cannot be certain of the total effect of poor administration on individual performance. The counselor is compelled to use his judgment in evaluating the influence of what he suspects and observes regarding administration. More research reveals that since conditions and methods under which tests are administered are of little consequence in *all* cases, the counselor will have to take into account the factors that *might* have negative implications for test performance and the interpretation of individual scores.⁵⁶ Some of the factors related to test administration, with their implications, are discussed in the following sections.

ACQUAINTANCE WITH TEST TO BE ADMINISTERED

Before a counselor gives a test he must be well acquainted with it. This acquaintance may be gained during the process of reviewing tests for possible use. Ideally, it would be desirable that the counselor take the test and have some practice in giving it. While the counselor probably cannot put himself completely in the position of the student who takes the test, he can listen for confusing or ambiguous directions and types of items. This process will give him some idea of what the student will meet and some of the questions he may ask. Even here, of course, the counselor will need to keep in mind that some sophistication on his part is assumed and that what *he* understands may not necessarily be understood by the student. By giving the test to those who will assist in the administration, the counselor will have an opportunity to become acquainted with details that he cannot take for granted. Regardless of how sophisticated he may be about testing, he cannot afford to

⁵⁶ For an excellent example of treatment of "Preparation for Testing" see the *Manual for the School and College Ability Test*, Princeton, N.J. Educational Testing Service, 1955. In this manual such items as "General Instructions," "Scheduling Tests," "Seating Arrangements," "Proctors," "Arranging Test Material," "Managing the Materials in the Testing Sessions," and "Role of the Examiner in the Testing Session," as they relate to this test, are discussed in detail.

make assumptions about his competence. Experience in giving one test does not assure proper administration of another. Test directions may be full of surprises and students may ask questions about a test that require good answers.

While the counselor in a given school will be limited by the facilities that are available, an attempt should be made to administer tests under the most ideal physical conditions. Suggestions for physical arrangements have been made in most texts on measurement, and those of Thorndike appear generally consistent with others. He suggests the following as characteristic of an ideal room for the administration of group tests.

1. It is quiet and free from disturbance of other activities.
2. It is well lighted and ventilated.
3. It provides each subject with a comfortable seat and a good writing space, preferably a desk or table.
4. It has appropriate size and shape and has sufficiently good acoustics, so that each person being tested can both see and hear the test administrator without difficulty.
5. It provides space so that test proctors can reach any subject being tested, to answer questions and to inspect his work.
6. It provides enough separation between testees to make cheating difficult or impossible.⁴⁷

In many high schools the school libraries come nearest to meeting the above requirements because they are equipped with tables. Since most current tests require use of separate answer sheets, sufficient space for them and the test booklet and comfortable arm support are desirable. The use of chairs with writing arms does not meet this need.⁴⁸

The counselor should give some thought to the seating of students. There may be a tendency for certain noisy groups of students

⁴⁷ Robert L. Thorndike, *Personnel Selection: Test and Measurement Techniques*. New York: John Wiley & Sons, 1949, p. 262.

⁴⁸ Arthur E. Traxler and Robert N. Hilkert, "Effect of Type of Desk on the Results of Machine-scored Tests," *School and Society*, September 26, 1942, 56:227-229.

to sit together if location is entirely optional and they may become centers of general distraction.

One of the assumptions that appears to be made in the administration and interpretation of group tests is that all students will have been equally interested and motivated while taking the test. The absence of discussion of motivation in test manuals suggests that the authors have concluded that high motivation is automatically obtained in group-testing situations. This is not necessarily the case. Students are frequently asked to take tests without being given any idea as to why they are taking them and how the results are to be used. They should know the purpose of testing and how the results will affect them. It is not improbable that many students have learned to be highly suspicious of tests.

Ordinarily, motivation for taking tests should stem directly from the counseling process. This precludes, of course, the kind of counseling that starts with the administration of tests. It does assume that tests will be used *only when, in the process of counseling, it is believed that test results will aid the student with his problems, decisions, and plans.* Perhaps the most desirable arrangements for testing would be found when, having reached a point in counseling where both the counselor and counselee see the need for further data, the student could be tested "on the spot" by the counselor or referred immediately to a psychometrist. Such arrangements are possible in some clinics, but staff and time are not usually available in most high schools. Until such arrangements are possible, the counselor may resort to group testing. He may keep a list of students who are to take tests until he has enough for a group administration.

Since motivation is so important, it would seem that test authors and publishers would do more about it. Most test manuals offer little in terms of prepared statements to be read to the subjects. Some of the examples following are typical of the perfunctory manner in which students are given tests. "As soon as booklets and

pencils are distributed, say: 'Fill in the blanks on the cover, but do not open the booklets.' Allow about two minutes, say: 'This is a test to see what you can do with your hands and eyes. Use the pencils provided' . . ."⁵⁹ The printed directions do not call for any statements about the purposes of taking the test or the uses to which the results will be put. In the opening paragraph of the manual it is stated unequivocally that "this battery of seven subtests provides objective measurement of the aptitudes which underlie successful performance of a wide variety of jobs of a mechanical nature." If this were true the students should be made aware of this fact, and how the results will relate to their plans. But since one of the principles of test administration is that prepared directions must be adhered to completely, it is assumed that persons giving the test will not go beyond printed directions.

The introduction of another test is similarly matter-of-fact: "To administer . . . address the pupils as follows: 'We are now going to give you some tests that measure your ability to think. I will pass out the test papers' . . ."⁶⁰ With the awareness that most students might have about the implications of the word *think*, such directions probably make many students apprehensive. Here again, as with the example above, the manual offers a list of six purposes for which mental tests are given but the word *think* does not appear.

Not all test manuals are as abrupt as those given above. One test manual advises the test administrator to: "Try to put the students at ease by explaining briefly why the tests are being administered. Stress the personal value of the tests for each student, so that the pupils will not only accept them but also put forth their best efforts. Before each test, state in a few simple words what the particular test is about; but avoid overdoing this phase so as not

⁵⁹ T. W. MacQuarrie, *Manual of Directions, MacQuarrie Test for Mechanical Ability*. Los Angeles: California Test Bureau, 1925, p. 6.

⁶⁰ *Manual for Otis Quick Scoring Mental Ability Tests*, Gamma EM, Yonkers, N.Y.: World Book Co., 1954, p. 2.

to encroach on the time allotted for testing.”⁶¹ This does not, however, tell the students why they are taking tests and what effect the results will have on their current progress, activities, or their futures. Since most test manuals offer little assistance on this matter the counselor must resort to his own devices.

In order to increase motivation for one group of students the following statement was distributed to them as they entered the testing room and was supplemented with similar oral comments.

TO PUPILS INCLUDED IN THE GUIDANCE STUDY

The tests and interviews which you are going to take will help us to help you to find out the things that you can do best. Because of the competition which exists in the world of work today, it is important for you to find what your abilities are in order to develop them to your best possible advantage. As a result of all these tests and interviews we hope to be able to advise you about various kinds of work and study. It is also our purpose to aid you in learning more about your own strong points and to help you to make the best of your opportunities. We hope that you will do your best on the tests which are given to you. Remember that they have nothing to do with your school marks. You may now ask any questions about the work.⁶²

Such orientation or motivational procedures will not, of course, guarantee best efforts of all and will not compensate for *other* factors that may cause individuals to do less than their best.⁶³ They

⁶¹ *Manual for Differential Aptitude Tests*. New York: The Psychological Corporation, 1947, pp. 3-4.

⁶² John W. M. Rothney. *Guidance Practices and Results*. New York: Harper & Bros., 1958.

⁶³ Another approach to this problem that appears to have many interesting possibilities is the use of a recently published pamphlet by Herschel T. Manuel. *Taking a Test. How to Do Your Best*. Yonkers, N.Y.: World Book Co., 1956. This pamphlet is designed to give students confidence and skill in taking tests, or more specifically, to give the students (1) a chance to learn what tests are for, what kinds of tests there are, how they are built, how the results are expressed, and what they mean; (2) an opportunity to learn good practices in taking tests; (3) actual experience with test materials (p. 3). In many ways the use of such a pamphlet to orient students could be justified, especially in terms of trying to equalize the advantage the testwise student may have over one who has little or no experience with tests.

may, however, remove some of the doubts and apprehensions that some students have when they are taking tests.

The sequence in which tests are presented may influence testees' efforts. If more than one test is planned, either in a single session or in close sequence, the one that is of most interest to the students should be given first. It is recognized, of course, that no test will arouse universal interest and the counselor will use his judgment after he has considered the nature of his group. The approach used in the Wisconsin Counseling Study in administering the tests of a differential series appeared to work well. Since circumstances required group testing, all the students participating in the study were grouped on the basis of need as determined by inspection of individual records. Those who planned to go to college or other kind of training after high school were given the Differential Aptitude verbal reasoning, numerical ability, and language usage tests.⁴⁴ Those who planned to enter mechanical, farming, and related fields were given the mechanical reasoning, numerical ability, and space relations subtests. Finally, those who intended to enter clerical and related fields were given the numerical ability, clerical, and language usage subtests.⁴⁵ In each case the students were told why they were grouped as they were. Students who took the mechanical sequence were administered the mechanical reasoning test first, the training group began with the verbal reasoning test, and the clerical group took the clerical aptitude subtest at the beginning of their testing period.

While it is desirable that the counselor administer all the tests to be given to his students so that he may observe reactions to tests at first hand, it will not always be possible. If he cannot do so he will do well to take exception to the statements found in many

⁴⁴ George K. Bennett, Harold G. Seashore, Alexander G. Wesman, *Differential Aptitude Tests*, New York: The Psychological Corporation, 1947.

⁴⁵ While it is not especially relevant to this discussion, it should be mentioned that other tests of the battery were given to students in any group on the basis of student requests and needs after the initial results were known and interpreted to the students.

test manuals, which suggest that virtually anyone who can read a manual can give tests. While statements of this kind may enhance sales by implying that a "specialist" is not needed, they take too many things for granted. They may assume, for example, that the personality and attitude of the test administrator and the rapport he has established with pupils in the general school situation are unimportant. Those who view the process of testing as essentially mechanical are not likely to inspire students to do their best. Those who have certain disciplinary relationships with students may not be accepted by all the testees. Those who resort to veiled threats and expressions of dire consequences or who may suggest that test performance is a life or death matter are not likely to do the best job. It is no doubt true that in any group situation there must be a certain amount of administrative control; it would seem equally true that the rapport in a group-testing situation should approach that of a one-to-one counseling situation as much as possible. As suggested earlier in this section, the counselor uses tests because they attempt to measure something that may have implications for a counselee's future performance. The performance sampled by the test presumably bears some relationship to future activity of interest or concern to the individual. One of the assumptions too frequently made when interpreting test scores is that the student knew what he was to do and did it as best he could. This may or may not be the case.

One important principle of testing requires that the directions provided in a test manual must be followed exactly as written. The reason for this is obvious, for the printed directions represent a part of the conditions and circumstances under which the test was standardized, the reliabilities were determined, the validity computed, and the norms constructed. It follows that if the data on these factors as provided in the test manual are to be used, the scores must be obtained in the same manner as those upon which the data in the manual are based.

In selecting a group test, initially, the counselor should choose one that has directions most likely to be understood by the "poorest" of his group. He will look for tests that have concise and clearly stated directions. He will look for tests for which the directions are not made too complex by the need to explain novel and involved physical manipulations. The directions should include suitable examples of the various types of items found in the test and provision should be made for adequate practice on such items. The directions of the test should anticipate as many of the students' questions as possible (Should we guess? Can we use scratch paper? How much time do we have? Can we check answers if we finish before time is up? How do we change an answer? etc.), and provide specific instructions for answering them.⁶⁶

Many tests do have inadequate or confusing directions and one example will suffice to illustrate the point. The following directions for one subtest of a mechanical ability test are read to students as they read along in the test booklet.

This is the practice page for the LOCATION test. Notice the letters in the large square, and the five dots in each of the small squares below. For each dot in a small square, there is a letter in the same place in the large square. When the examiner says GO, but not before, put right on each dot the letter that stands in its place in the large square. For instance, the upper dot in the first small square is in the position of the letter K in the large square, so you will put a letter K on that dot. READY, GO! (THIRTY SECONDS) READY, STOP! In the small square at the left you should have V, K, N, E, K. In the one at the right you should have U, E, M, O, C. (Take a little time here for consideration of errors.) Turn to page 11.⁶⁷

⁶⁶ One of the authors recalls his experience when, as an inductee in World War II, he was given the Army General Classification Test. When the group was invited to ask questions regarding the directions, he asked whether there was a penalty for omissions. The response to his inquiry was, "Are there any other questions?" Not many months later, when assigned to give the tests, he learned that there was no scoring penalty for omissions and had another form administered to him. With his question answered, he improved his original score by 20 points.

⁶⁷ *Manual of Directions. MacQuarrie Test for Mechanical Ability.* Los Angeles: California Test Bureau, 1953, p. 8.

Directions such as these would seem so complex as to confuse even the test sophisticate. If the counselor must use tests with directions as difficult as those in the example, he must be alert to the possibilities that students will misunderstand them.

The counselor, however careful and demanding he may be in making test selections, will not find any test with directions that will be understood by all his subjects. The range of individual differences within a group being tested, the previous experience in taking tests, the variables such as reading skills that might relate to responding to directions, the "set" about tests that may cause an individual to misinterpret test directions make this obvious. The fact that most students will be able to follow the directions as provided in the manual does not relieve the counselor of being alert to the exception.

The counselor, then, will watch during the administration of a test for students who seem puzzled, who look around to see what others are doing. He will walk unobtrusively around the test room after the students have started to check on the students' approach and progress. He will question scores obtained from papers when only a few items are completed and question those papers where the initial items are missed. He will always keep in mind that the usual invitation for questions included in directions may not bring out those in the minds of the shy students.

The clinician who administers individual tests can observe the behavior of subjects during the testing situation and frequently these observations are of more help than the test scores in understanding the individual. Many of the discussions of observation during testing emphasize the value of such observations.⁶⁸ It is possible that they may also have value in the interpretation or significance of particular scores.

It is important that test behavior be observed and any behavior

⁶⁸ Ruth Strang, *Counseling Techniques in College and Secondary School*. New York: Harper & Bros., 1949, p. 52. See also the discussion on this point in Donald E. Super, *Appraising Vocational Fitness*. New York: Harper & Bros., 1949, pp. 81-84.

that might influence the test score be noted on the cumulative or other test record. As a further check on test behavior and performance, it is good practice for the counselor to elicit the student's reaction to the test situation when he is interpreting the results to the student. While some students may attempt to rationalize a poor or mediocre score, others may reveal some of the conditions of testing that were not observed—or observable. In general in such cases it is better to schedule a retest than to refuse to give the student the benefit of any doubt.

SCORING OF TESTS

Speed, accuracy, and economy of time and effort are major considerations involved in any discussion of test-scoring procedures. Publishers' brochures often capitalize on this concern for the more "practical" mechanics of test use and promote various "gadgets," self-scoring devices, "Scoreeze" formats, machine-scoring answer sheets, transparent plastic templates, and other streamlined innovations to make scoring seem easy and rapid. The choice of a test is often determined primarily by the advertised simplicity of scoring.

Counselors are interested in any time saved in the scoring of tests since every hour saved in this operation makes it possible to spend more time in interviewing students. But if the counselor is to approach interviews with confidence that his test scores are accurate, he must pay close attention to the problems involved in scoring.

In many large school systems special test-scoring staffs are assembled and trained for the arduous task of test scoring. Other schools use professional test-scoring services of an agency such as the Educational Records Bureau or take advantage of the test-scoring services offered by several test-publishing companies.⁶⁹

⁶⁹ Arthur E. Traxler, "Administering and Scoring the Objective Test," in E. P. Lindquist (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 34-35.

While the cost may be small considering the amount of time and money already invested in the testing program, many schools do not provide for it in their budgets.

When teachers are to score the tests, an in-service training program about the objectives of the testing program, best ways to interpret test scores, and careful presentation of the seemingly simple step-by-step scoring procedures themselves are required.⁷⁰ Attention needs to be given to accuracy and provision should be made for every paper to be scored at least twice in order to spot clerical errors.

Some schools that do not have a special test-scoring staff have solved their problems by appointing a teacher from each department or instructional level who is responsible for the administration and scoring of all standardized tests. Under this plan it may be necessary to hire substitutes for these teachers while they are performing such duties. The advantage in this procedure is that training, practice, and the development of accurate techniques in scoring are more apt to come from persons who are given this specialized responsibility than when all teachers in the school are required to participate in the scoring program.

USE OF SEPARATE ANSWER SHEETS

Many tests provide a separate answer sheet for the students' responses. While this simplified, mechanical aid seems to have distinct advantages over the older procedure of marking the correct answers on the test booklet, it is not without some disadvantages. The first of these is the effect upon the student. When separate sheets are used with tests whose norms are based on the booklet marking and scoring typical of most tests published more than ten years ago, some caution in interpretation is necessary. The extra

⁷⁰ A good plan for the operation of a test-scoring unit is given in Arthur E. Traxler and Others, *Introduction to Testing and the Use of Test Results in Public Schools*, New York: Harper & Bros., 1953. See Chapter 6, "How Should Tests Be Scored?" especially pp. 37-42.

perceptual and manual skill called for in the manipulations of separate answer sheets may bring incorrect inferences about the typical performance of some students. This is doubly true with highly speeded tests where precious seconds may be lost in the handling of the answer sheet.

The use of separate answer sheets imposes some rather rigid limitations on the type of test item that may be used. They usually require a multiple-choice or true-false response in order to fit the structural limitations of the sheets. The counselor should be aware of this limitation in his inferences about performances of the counselee, for few of a student's real problems appear in neat rows of five alternate responses.⁷¹

One distinct disadvantage of the separate answer sheet, scored by an agency, is found in the extra tasks involved if the counselor wishes to make a study of responses to particular items.⁷² On the older type of booklet-marked achievement test it was possible to make a diagnostic summary of particular items missed by certain pupils. While there is nothing to prevent a counselor's requesting an item analysis from a scoring-machine operator, extra interest and motivation for child study seem to be called for if he is to go beyond the total score and use the test diagnostically. In some high schools where separate answer sheets are machine-scored and whole classes of students are given multiple aptitude test batteries, only the profile sheets that contain scores of his counselees are handed on to the counselor. He has very little information about the specific items that his counselee missed or answered correctly.

Manual scoring of test booklets is common for individual intelligence tests and for several of the older aptitude and intelligence tests. Scoring procedures for the older forms of such tests as the

⁷¹ Arthur E. Traxler, "Administering and Scoring the Objective Test," in E. F. Lindquist (Ed.), *Educational Measurement*, Washington, D.C.: American Council on Education, 1951, pp. 384-388.

⁷² Gerald M. Rapaport and Irwin A. Berg, "Response Sets on a Multiple-Choice Test," *Educational and Psychological Measurement*, Spring, 1955, 15:58-62.

Otis, Minnesota Paper Form Board, and the Minnesota Clerical tests are done by use of fan, strip, or cutout stencils. This procedure, which allows for some diagnostic study of the counselee's test responses, may still be generally preferred over a separate answer sheet.

MECHANICAL CONSIDERATIONS

Such factors as typography of the test materials, reading ease of the test items, and complexity of the actual physical handling of the testing materials by the testee must be considered carefully by the counselor. Since there is very little research on these matters, he is forced to rely on his own judgment about their effects.

Comparison of the typography of older tests with more recently released tests reveals that few changes have been made. The amount of white space around test items, size of type, and general appearance of such older tests as the National Intelligence Tests, published in 1920, compare favorably with tests still commonly used. Some departure from standard presentation has been attempted by several test publishers. The California Tests of Mental Maturity are printed with green ink on white paper in an attempt to improve ease of reading. The recently published School and College Ability Tests of the Educational Testing Service alternate red with black ink in an effort to improve the typography of their test. Few attempts have been made to evaluate the effects of these innovations on the performance of testees. Without such data the counselor must be careful in assessing the degree of improvement that these typographical variations introduce.

In some of the older tests the items are crowded together too closely and seem to impose some disadvantages for those students with less than normal visual acuity.⁷³ The older Henmon-Nelson

⁷³ Fritz Forbes and William Cottle. "A New Method for Determining Readability of Standardized Tests." *Journal of Applied Psychology*, June, 1953, 37:185-90; John Pierce-Jones. "The Readability of Certain Standard Tests." *California Journal of Educational Research*, March, 1954, 5:80-82.

Tests of Mental Ability, the earlier editions of the California Tests of Mental Maturity, the Van Wageningen Reading Readiness Test, or the Myers-Ruch High School Progress Test, all published prior to 1940, serve as examples of tests whose typography would seem to leave something to be desired.

Several research studies indicate that some individualized tests require greater reading skill on the part of the testee than he is apt to have developed at the time of testing.

The format of most modern standardized paper and pencil tests that use separate or machine answer sheets is straightforward and simple. The testee is given an answer sheet, a test booklet, and possibly an electrographic pencil. The directions are given on the test booklet for various subtest tasks. The counselor reads the directions to the subjects and may illustrate the correct marking procedures on the blackboard. This seems to minimize errors in the physical aspects of taking the test. Despite the seeming simplicity of this task, several kinds of errors may still be introduced.

If a testee places his test booklet for the Clerical subtest of the Differential Aptitude Tests to the side and his answer sheet directly in front of him he is penalized because he uses excessive eye movements as he goes from one to the other and tries to keep his place on the complex answer sheet. The use of a paper straight edge, as some testees discover, makes this visual task much simpler. The Minnesota Paper Form Board is folded so that a hurried testee may miss items 17-48 without being aware that he has done so. Care must be taken with the SRA Primary Mental Abilities Tests (or any of the Science Research stepped-down test booklets) that the answer sheet is aligned accurately with the proper item in the test booklet or a whole subtest may be marked incorrectly.

Only prolonged familiarity with and use of testing materials will provide the counselor with an adequate frame of reference with which to judge most of the mechanical aspects of tests. He must be constantly alert in his use of testing materials to evaluate

the possible effects of the mechanical make-up of the test itself as he attempts to interpret scores to his counselees.

SUMMARY

In this chapter several criteria that the counselor must keep in mind when selecting tests have been presented. It has been pointed out that while many authors of tests make broad claims for usefulness and application in many situations, many of these claims are more implied than demonstrated. It has been suggested that the counselor must demand *evidence* that the tests accomplish what is claimed for them. It has been further pointed out that, since much of the counselor's work involves predictions of some future performance with varying time lapses, it is essential that reliability data must be supplemented by some evidence of stability.

It has been shown that the separate answer sheet method is not clearly superior to the hand-scored answer sheet in saving time or providing useful data for counseling. When the answer sheets are checked visually for careless marking, the number of answer sheets that can be handled in an hour by a trained manual scorer approximates the number that can be processed by a machine-scoring unit. Careful analysis of costs, availability, and extra supplies should be determined before the scoring machine is accepted as a solution to the test-scoring problem.

It has been suggested that the counselor must look for test data that will describe the size of the norm group with which a counselee is to be compared as well as the psychological, socioeconomic, geographic, and educational characteristics of the group. He will also seek data that will permit him to make comparisons of counselees' scores with those who have entered training of the kind the student plans to undertake. The counselor will find, more often than not, that adequate data of this kind are not available in test manuals.

Attention must be paid in selection tests to such factors as administration, scoring, and mechanical features of the instruments. While these are not as vital as those of validity, reliability, and norms, they can seriously affect the scores obtained. The counselor will not be misled in his selection of tests by extravagant claims of ease of administration and speed of scoring but will seek those tests whose usefulness in counseling individuals is backed by adequate evidence.

DISCUSSION QUESTIONS AND EXERCISES

1. Select several representative tests of mechanical (or other) aptitude and construct a list comparing them on the basis of criteria used in validation. To what degree do these criteria afford a logical basis for prediction? What other criteria might have been used in the area of testing you selected? How dependable or reliable are the criteria used? Assuming that a shop foreman may not rate a worker's performance in the same way twice, how useful are such ratings likely to be as a validity criterion? Would this apply also to teacher's grades?
2. As a class or group project, review 25 tests of various kinds and classify validity data under the headings *content*, *predictive*, *concurrent*, and *construct*. For individual counseling which is most useful? How many tests offer data on more than one type of validity? In terms of the stated purposes of each test, comment on the appropriateness of the type of validity data used.
3. Below is a distribution of the percentiles achieved by 731 students. They took the Henmon-Nelson Test of Mental Ability while they were in the tenth grade and a year later in the eleventh grade. The reliability coefficients (test-retest) reported in the manual for the test were .900 and .887. The tests were administered by members of the faculties of the high schools.
 - a. What factors might have influenced the variability in test performance indicated in the table?
 - b. Could the coefficient of correlation between the two sets of

TABLE 13. First Administration (Tenth Grade)

Percentiles	0	11	21	31	41	51	61	71	81	91	Total
	10	20	30	40	50	60	70	80	90	100	
91-100						2	2	2	24	53	83
81-90					3	4	11	12	26	21	77
71-80		1			3	6	17	15	21	11	74
61-70			3	7	6	14	17	9	9	1	66
51-60		1	2	11	9	10	12	11	5		61
41-50	1	1	11	16	11	15	16	8			79
31-40	6	10	10	11	21	14	16	1			89
21-30	4	9	9	10	8	5	3	3		1	52
11-20	17	21	14	12	3	2	1	1			71
0-10	49	14	8	3	1	3			1		79
Total	77	57	57	70	65	75	95	62	86	87	731

scores be considered as a coefficient of test-retest reliability? A measure of the stability of test scores?

- c. In what percentage of the cases do you believe that the *variability of performance from one year to the other* might influence the counseling of students?
 - d. Does the variability indicate that the test should be given twice before the results can be used with confidence?
 - e. In view of the distribution obtained, of what value are the reliability coefficients reported in the test manual?
 - f. Of what significance is the fact that subjects at the extremes seem to be more consistent than those in the middle ranges? Would you attribute it largely to variability of individuals or the nature of percentiles?
 - g. Using the standard errors of measurement given in the test manual and the table of percentiles, compute the possible variability of the subjects in the diagonals who seem to be highly consistent.
4. Select a test for which the standard error of measurement is offered in the test manual. Plot the error in points against the scores on the percentile table. What range of scores is represented in plus and minus one standard error from the score equivalent to the fiftieth percentile? What range of scores, and hence percentile rank, is noted when two standard errors are plotted. Three? On

the basis of these data, what are the implications for the interpretation of test scores?

5. The following array of scores was obtained from the administration of the Differential Aptitude Verbal Reasoning Test to ninth grade boys in public high school:

11, 1, 9, 24, 17, 7, 27, 2, 10, 13, 10, 8, 13, 8, 13, 8, 8, 9, 8, 18, 18, 23, 6, 25, 12, 10, 8, 32, 16, 20, 15, 19, 12, 34, 13, 10, 10, 12, 26, 4, 23, 23, 27, 6, 1, 29, 20, 11, 24, 7, 17, 4, 17, 22, 6, 18, 18, 3, 2, 6, 29, 25, 20, 4, 28, 21, 4, 15, 4, 14, 16, 11, 3, 0, 5, 15, 20, 8, 14, 12, 23, 31, 24, 33, 6, 27, 14, 11, 18, 22, 21, 3, 11, 5, 15, 11, 5, 12, 6, 17, 22, 24, 17, 22, 21, 15, 10, 16, 13, 12, 26, 18, 8, 4, 40, 14, 16, 19, 23, 13, 21, 13, 7, 8, 30, 22, 11, 14, 6, 16, 17, 24, 9, 9, 8, 5.

Using procedures suggested in any standardized text in statistics, construct an ogive curve for this array of scores. Make a percentile table, comparing the percentile equivalents obtained above and from those published in the manual for the Differential Aptitude Test Battery. What differences do you find in the comparison? What factors might account for the differences or similarities? What are the implications for counseling an individual student in terms of *his* status in *your* school?

6. Discuss the relative merits of national and local norms. Under what circumstances or in what counseling situations would each be most appropriate?
7. Select a standardized psychological test and note the directions to be used in giving the test. What are the possible misinterpretations of directions on the part of the student? How would you rewrite the directions to take care of possible misinterpretations of the task to be done? Under what conditions of use might the directions be rewritten?
8. For the test selected in Exercise 7 above, prepare what you believe would be a statement that might be used to motivate the students to do their best.

REFERENCES

American Psychological Association. *Technical Recommendations for*

- Psychological Tests and Diagnostic Techniques*. Supplement to *Psychological Bulletin*, March, 1954, 51. Washington, D.C.: American Psychological Association, 1954.
- Anastasi, Anne. *Psychological Testing*. New York: Macmillan, 1954.
- Anastasi, Anne, and Drake, J. "An Empirical Comparison of Certain Techniques for Estimating the Reliability of Speeded Tests." *Educational and Psychological Measurement*, Autumn, 1954, 14:529-540.
- Bittner, R. H., and Wilder, C. E. "Expectancy Tables: A Method of Interpreting Correlation Coefficients." *Journal of Experimental Education*, March, 1946, 14:245-52.
- Brown, Clarence W., and Ghiselli, Edwin E. "Some Generalizations Concerning the Validity of Aptitude Tests." *Personnel Psychology*, Summer, 1953, 6:139-50.
- Cottle, William. "A Form for Evaluating Standardized Tests." *Occupations*. December, 1951, 30:188-194.
- Cronbach, Lee J. *Essentials of Psychological Testing*. New York: Harper, 1949.
- Cronbach, Lee J. *A Consideration of Information Theory and Utility Theory as Tools for Psychometric Problems*. Champaign, Illinois: Bureau of Research and Service, University of Illinois. November, 1953.
- Cronbach, Lee J., and Meehl, Paul E. "Construct Validity in Psychological Tests." *Psychological Bulletin*, July, 1955, 52:281-302.
- Cureton, Edward E. "Validity, Reliability, and Baloney." *Educational and Psychological Measurement*, Spring, 1950, 10:94-96.
- Cureton, Edward E. "Validity" in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 621-694.
- Flanagan, J. C. "Units, Scores and Norms" in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 695-763.
- Gaylord, R. H., and Stunkel, E. R. "Validity and the Criterion." *Educational and Psychological Measurement*, Summer, 1954, 14:294-300.

- Jenkins, J. G. "Validity for What?" *Journal of Consulting Psychology*, March-April, 1946, 10:93-98.
- Johnson, Ralph H., and Bond, Guy. "Reading Ease of Commonly Used Tests." *Journal of Applied Psychology*, October, 1950, 34:319-324.
- Jones, Arthur J. *Principles of Guidance*. Third Edition. New York: McGraw-Hill, 1951.
- Lennon, Roger. "The Test Manual as a Medium of Communication." *Proceedings of the 1953 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1954.
- Lorge, Irving. "The Fundamental Nature of Measurement," in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 533-559.
- Manuel, Herschel T. *Taking a Test. How to Do Your Best*. Yonkers, N.Y.: World Book, 1956.
- Patterson, C. H. "The Interpretation of the Standard Error of Measurement." *Journal of Experimental Education*, March, 1955, 23:247-254.
- Pollack, Abraham B. "How to Tell Whether Aptitude Tests are Trustworthy." *Business Education World*, December, 1949, pp. 170-172.
- Remmers, H. H., and Whisler, L. "Test Reliability as a Function of Method of Computation." *Journal of Educational Psychology*, February, 1938, 29:81-92.
- Rothney, John W. M., and Roens, Bert A. *Counseling the Individual Student*. New York: Dryden Press, 1949.
- Selby, P. O. "Are Predictive Tests Reliable?" *Journal of Business Education*, October, 1941, pp. 13-15.
- Stuit, D. B. "The Preparation of a Test Manual." *The American Psychologist*, May, 1951, 6:167-70.
- Super, Donald E. *Appraising Vocational Fitness*. New York: Harper, 1949.
- Thorndike, Robert L. *Personnel Selection*. New York: Wiley, 1949.
- Thorndike, Robert L. "Reliability" in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951, pp. 560-620.

- Travers, Robert M. W. *Educational Measurement*. New York: Macmillan, 1955.
- Traxler, Arthur E. "Administering and Scoring the Objective Test." in Lindquist, E. F. (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.
- Traxler, Arthur E., and Others. *Introduction to Testing and the Use of Test Results in Public Schools*. New York: Harper, 1953.

CHAPTER IV

Test Scores: Etiology and Interpretation

When a counselor notes test scores similar to those below he

TABLE 14. Test Scores

Date	Test	Form	Score	Derived Score
1/26/58	Jones Test of Mental Ability	A	42	IQ 100
3/14/58	Iowa Silent Reading (Median Standard Score)	A	144	grade 7-1
4/20/58	Bennett Mechanical Comprehension	A	45	75th percentile

may feel that he can carry on from there—that he has an indication of his counslee's performance as specified by the title of the tests in terms of quality and quantity. If so, he may be tempted to make such statements as the following:

"He has an IQ of 100."

"His reading ability is at the seventh grade level."

"His score on a mechanical aptitude test is at the 75th percentile."

These are samples of the kinds of statements commonly made by teachers and counselors when they are reporting about pupils or counselees. From such statements it is not uncommon for them

to make deductions about whether the student is working up to capacity or even beyond his capacity, to make recommendations about choices that must be made, or to predict his performances in future educational or vocational activities. They should realize, however, that the deductions that can be made, the counsel that can be offered, and the actions that can be taken on the basis of the test scores are determined by many factors that lie behind and beyond them. If the counselor is going to use test scores wisely he will find that he must proceed in opposite directions: back into the test and forward into the implications of scores. His findings on the former will determine how far, if at all, he can extend the latter. In this chapter the factors that are inherent in test scores, factors which may influence them, and the implications of these factors in their interpretation will be examined.

SELECTION OF A TEST

The reader will have discovered through his review of test publishers' catalogues, specimen sets of tests, and other sources such as the *Mental Measurement Yearbooks*, that the counselor will have many hundreds of tests from which to choose. It is essential that the counselor will be guided in his choice of tests by the criteria presented in this and the previous chapter. Even with the most careful initial selection, however, he will find that he is dealing with imperfect instruments at best, and that his "best" will have its own unique characteristics and properties. He will find that each test is different even though it may carry a title similar to some other test, and that there is no "standard" interpretation that can be applied even to tests which purport to measure the same thing. Each test must be interpreted in light of *its* background and development, and each test may be interpreted differently for different individuals according to the factors or circumstances which surround the individual's approach to the test and *his* background. The counselor, then, will ask many questions about a test

prior to its selection and, having selected an instrument, he will keep many of the same questions in mind when interpreting the results.

As the counselor contemplates tests-result entries such as the above, then, he must ask himself what factors need to be considered before such entries become really meaningful. What does each entry indicate about the student on whose record it appears? What are the implications of the scores and what can they possibly suggest about next steps in working with the counselee? The answers to such questions will require a very thorough investigation of the competency of the test authors and publishers, the assumptions they have made in the construction, scoring, and norming of the test, the quantity and quality of the evidence they have presented concerning the efficiency of the instrument, the conditions under which the test was given, the response of the student to the test situation, and the care with which the scoring and tabulating of scores was done. These are all factors influencing a given test score, and until the counselor has become convinced as the result of his investigations that none of these circumstances are faulty, he cannot use the test score with assurance. The inquiries that he should make and the kinds of information he is likely to find are described in the following pages.

Before proceeding with a discussion of the factors that may influence test scores, it should be pointed out again that scores obtained from tests with similar names often yield quite different results. It is always desirable, therefore, to give the specific name of the test from which a score was derived whenever it is mentioned. It is always necessary, too, to remember that no one *has* an IQ or an ability. The test score can indicate only his performance on a given test at a given time, and cannot possibly indicate *possession* of an IQ, a percentile, a grade level, or an aptitude. Even when we use these cautions and say, for example, that a counselee has scored at a certain level on a certain test at a particular time, it is possible to make only certain limited kinds of

interpretations since the size and significance of the score will have been influenced by many factors.

FACTORS INFLUENCING TEST SCORES

AUTHORS OF TESTS

To suggest that the author of a test can influence a test score seems obvious, but the point here is not always obvious to the test user. While there are many tests with essentially the same title, "(such-and-such) Test of Mental Ability," the underlying concepts of mental ability and procedures for measuring it may vary greatly. They depend upon the training, psychological orientation, and concepts of intelligence held by their authors. These may be determined by their adherence to a particular school of psychological thought, their philosophy, and even their sociological and statistical orientations. When one buys any author's test, he also buys and uses the author's assumptions and concepts. These may be crucial, but they may very well be overlooked by the counselor when he interprets a score from a "mental ability test." Definitions of terms would not be a factor if we had universal agreement as to what mental ability really is, how it is manifested, and how it can be measured, but since this is not the case, the test author is very much a part of the test score.

Counselors may begin the study of a test by examining biographical sketches of test authors provided in a number of professional publications.¹ From them, he may learn of the author's training, experience, and competence in the field of measurement. In some cases he will find biographical sketches of test authors in the manual which accompanies the test. Attempts should be made

¹ Jaques Cattell (Ed.), *American Men of Science*, Vol. III, *The Social and Behavioral Sciences*, New York: R. R. Bowker, 1956.

J. McKeen Cattell, Jaques Cattell, and E. E. Ross (Ed.), *Leaders in Education*, 2nd ed., Lancaster, Pa.: The Science Press, 1941.

Directory of the American Psychological Association. Washington, D.C.: The American Psychological Association, 1957.

to obtain information about the current as well as previous activities of the authors to determine whether or not they are carrying on professional rather than commercial activities. The counselor should also be aware that an author may have achieved high prestige for many years because he presented theories and practices that seemed to be sound at the time but that later proved to be very unsound.

Study of the author's record must be supplemented, however, by examination of his actual performance in the construction of the test under consideration. If the counselor finds, for example, that authors of unquestionable repute claim that the best evidence of the validity of the test they offer is to be found in its successful (undefined) use over a period of years, he must weigh current performance more heavily than previous reputation. If an author has been active in psychological and educational measurement over such a long period of time that his name has become almost a household word, but is still evasive in the discussion of norms and validity in a test manual, the latter performance is the one about which the counselor must be concerned. And if an author violates in his construction of a test what he proclaims to be essential in his books and pamphlets on measurement, the counselor must question the test, the writings, or both.

Perhaps the counselor who has noted the issues raised above has reached the conclusion that the reputation of an author can suggest general competence, but that it does not guarantee that each test he produces will be useful and dependable. Study of the biographical sketch of an author is a first essential step in the process of test selection, but it must be followed by intensive study of the information provided in the test manual.

PUBLISHERS OF TESTS

It is admittedly a difficult task for a counselor in a public school or college to make a judgment about something as intangible as

the reputation of a test-publishing house. In a field of publications that has become highly competitive and profitable in the last generation (over 75 million tests were used in schools in the past year), it has become increasingly necessary for test users to distinguish between sales promotion and careful research in test design and improvement. Counselors need some background of experience in dealing with the products of the several major test-publishing concerns before they can develop some basis of judgment of the "name" or reputation of a particular publisher. Some criteria that may help him in this evaluation have been presented in Chapter II.

Information about publishers of tests is difficult to obtain unless the counselor has had long experience in dealing with them. He must judge them by examining their catalogues to see if they continue to offer for sale tests that are demonstrably obsolete or inadequate, if they restrict the sale of test materials to competent persons, and if they provide adequate descriptions of their materials. In this connection, it is particularly important to note whether the test publisher makes the distinction between validity and reliability and does not imply that reliability means dependability. Some evidence of a publisher's position may be obtained from his willingness to answer, without equivocation, letters requesting supplementary information about his tests or asking for explanations about test data that are not clear. The quality of discussions about tests in the pamphlets that they issue may give some clues concerning the publishers' policies. The counselor will examine that material to see if they present information and research results that may be useful regardless of whether it concerns or requires the use of only the tests they produce. He will examine all the publishers' material, too, to see if they propose ready-made testing programs that may seem good to psychometrically naïve persons but do not take into consideration local circumstances and needs.

The reputation of the publisher becomes important when the

counselor is considering the final score obtained on the test, particularly if it suggests questionable ethics and practices. If the counselor is inclined to say, "They know more about tests than I do," or "No company can afford to market a questionable instrument," then he accepts everything about the test with which the publisher had anything to do. The fact that not all published tests are equally good, theoretically sound, and statistically acceptable, but still may find a publisher, suggests that not all publishers set for themselves the strict standards that should be met by all. Since there is a tremendous market for psychological tests, and since their publication in many instances has become a highly competitive profit-making venture, not all publishers have resisted the temptations (or even need) to meet the competition by marketing tests without sufficient development, standardization, or validation at the ultimate expense of the student on whose cumulative record the score is noted.

Occasionally a good company may publish a very poor test. The counselor must always, therefore, look carefully at each test produced by any publisher to see if it is likely to provide a meaningful score for the particular counselee with whom he is working.

The publisher's role in the production of a test may vary, of course. In some instances, publishers assume the responsibility for all the statistical development of a test, relying on the author only for basic test items and some other details. On the other hand, some publishers, with some tests, do little more than provide a "drafting" or "design" function and act as the marketing agent, doing nothing with statistics, development of norms, reliability, validity, or subsequent revision of a test.

TEST TITLES AND ASSUMPTIONS INVOLVED IN CONSTRUCTION OF TESTS

It was indicated in an earlier chapter that the title of a test does not indicate the kind of items it contains. Many tests are called

tests of *mental ability* or *mental maturity*, for example, but since there are many definitions of such terms it will be necessary for the counselor to determine which of them is employed by the authors of tests. This can be done by examining the writings in test manuals, in books, and journal articles. The following illustration of a method for doing so may be helpful to counselors who are in the process of selecting tests.

If one were to consider use of the Terman-McNemar Test of Mental Ability,² for example, he might examine the writings of the major author to see if he can get definite statements or clues concerning his concept of mental ability.

Terman was the author of the Stanford-Binet test. Since the Terman-McNemar Test of Mental Ability seems to be an attempt to put the principles used in the Stanford-Binet into group test form, it is probably safe to assume that the terms "mental ability" and "intelligence" are used interchangeably. Terman, the major author, is on record with respect to the definition of intelligence (or mental ability) in the following words:

In the case of intelligence it may truthfully be said that no adequate definition can possibly be framed which is not based primarily on the symptoms empirically brought to light by the test method. The best that can be done in advance of such data is to make tentative assumptions as to the probable nature of intelligence, and then to subject those assumptions to tests which will show their correctness or incorrectness. New hypothesis can then be framed for further trial, and thus gradually we shall be led to a conception of intelligence which will be meaningful and in harmony with ascertainable facts.³

Terman has attempted, as Binet did, to analyze some of the mental processes which the tests bring into play. The chief procedure is, as noted in the quotation, to base definitions primarily on the symptoms empirically brought to light by the test method.

² *Terman-McNemar Test of Mental Ability*. Yonkers, N.Y.: World Book Co., 1942.

³ Lewis M. Terman and Maud A. Merrill. *Measuring Intelligence*. Boston: Houghton Mifflin Co., 1937, p. 4.

The method here consists of "obtaining a general knowledge of the capacities of a subject by the sinking of shafts at critical points."

Using the "sinking of shafts" method, Terman described the procedure for selecting the items of the original group test of mental ability. The items were selected as follows: "The test as it now stands is composed of questions and problems which were selected from a much larger number by correlating each separate item with a dependable measure of mental maturity. The criterion used for this purpose was a composite which included grade location, age, total score on a two-hour mental test, and ratings of the pupils by from two to five teachers on intelligence and quality of school work. Try-out of these resulted in the elimination of three of the thirteen tests, and in the reduction of the 610 items in the remaining tests to 370. All items which failed to differentiate pupils of known brightness from known dullness were eliminated." ⁴

The revisions of the old form were designed to make the test now under consideration provide "more homogeneous material in order to have a test more highly saturated with a common ability." ⁵ The arithmetic and numerical subtests were eliminated so that the "scores of any two individuals are more nearly comparable qualitatively; i.e., they lie along the same continuum. This continuum may be characterized as *general verbal intelligence*. This particular change has, of course, been prompted by recent developments in factor analysis."

Some other minor revisions to permit more rapid scoring and some test substitutions have been made, but the reasons for the latter changes are not given in the manual. It is indicated that the correlation of the revised test with the original test is .91 "which

⁴ *Manual of Directions for the Terman Group Test of Mental Ability*. Yonkers, N.Y.: World Book Co., 1920.

⁵ *Manual for the Terman-McNemar Test of Mental Ability*. Yonkers, N.Y.: World Book Co., 1920.

indicates that the new test can be considered to be measuring essentially the same basic abilities covered by the original forms."

At this point the counselor who proposes to use the test must stop to consider some of the factors that are involved. The title of the test is the *Terman-McNemar Test of Mental Ability*, but when he reads the manual he finds that the test attempts to measure *general verbal intelligence*. Now he must decide whether he wants to use an instrument that excludes arithmetic and numerical subtests. Shall he accept the conclusions arrived at by factor analysis, or any other method that happens to be currently popular, that a test of *mental ability* may exclude any reference to performance with numerical symbols? Will it be clear when the score appears on the student's record that this test was limited to *verbal* materials? Will he now have to add tests that *do* sample spatial and numerical materials?

In the illustration given above, it has been rewarding to go back and find rather complete and clear statements about what the authors of the test were attempting to measure. Whether or not one agrees with the authors, one knows what the authors mean by the title of the test and the assumptions that they have employed.

The fact that all tests bearing similar titles are not necessarily similar and do not use the same basic assumptions is revealed when the statements given above are compared with those found in the manual of another test which also has the term "*mental ability*" as part of its title. "A measure of a pupil's brightness, called an Intelligence Quotient (IQ), is sometimes found by dividing the pupil's mental age by his chronological age. A measure of brightness of a pupil comparable to an intelligence quotient (IQ) obtained on the Binet Scale may be found by comparing his score in the Gamma Test with the norm for his age . . . A measure so found is not a quotient, but it is called an 'IQ' because it has the same significance." *

* *Manual for the Otis Quick-Scoring Mental Ability Test, Gamma Form EM.* Yonkers, N.Y.: World Book Co., 1953, p. 5.

At this point the counselor may discover that what he has is a "deviation" (from the norm) IQ rather than a quotient. He may also be led to believe, because the manual suggests that he has something "comparable to an intelligence quotient (IQ) obtained on the Binet scale," that there is no difference—that he has obtained in the 30 minutes required for the testing something that would have taken an hour or more to get if the individually given and highly respected Binet was administered.⁷ But is this indeed what he has? How "comparable" are the IQ's so obtained? Is the IQ exactly the same or only an approximation of the Binet? Are they interchangeable? Or are they just "comparable" in principle or theory? He gets some hint as he reads further: "Gamma IQ's found in this method tend to be somewhat less variable than ordinary IQ's, that is, they tend to be somewhat nearer to 100. This fact should be borne in mind if comparisons are made between Gamma IQ's found above, and ordinary IQ's found by the division method."⁸

Now he must ask, how much "less variable"? Does this mean he needs to compensate at the extremes—that the low IQ tends to be higher than it *should* be and that the high IQ tends to be lower than it should be? Does not this fact need to be borne in mind whether or not he happens to be comparing IQ's derived from different sources and by different methods? This variability, where the IQ is 100, may not be of consequence. A score at the extremes *may* be, but the counselor receives no help from the manual in terms of how much compensation is needed.⁹ In any event, if the

⁷ Anne Anastasi, *Psychological Testing*. New York: Macmillan Co., 1954, pp. 85-87.

⁸ *Manual for the Otis Quick-Scoring Ability Test*, *op. cit.*, p. 4.

⁹ The publishers of this test have supplied some of the data missing in the manual in their *Test Service Notebook*, No. 11, "A Comparison of the Results of Three Intelligence Tests," a report of an investigation by Roger T. Lennon. Psyche Cattell has also discussed this variability in relation to earlier forms of the Otis tests in two articles: "IQ's and the Otis Measure of Brightness," *Journal of Educational Research*, June, 1930, 22:31-35; and "Why the Otis 'IQ' Cannot be Equivalent to the Stanford-Binet IQ," *Journal of Educational Psychology*, November 1931, 22:599-603.

counselor is going to use the test results, he will need to be aware that not all IQ's are the same just because they might be so designated on the cumulative record, and that variability may be important in the case of the particular counselee with whom he is working.

The two illustrations given above have indicated that two tests with similar titles may differ in the basic assumptions used by the author and therefore in the results produced.

In his consideration of tests of mental ability the counselor might also want to examine a test with a similar title, but with somewhat different assumptions involved. He might then turn to a study of the California Test of Mental Maturity.¹⁰ This test was first introduced at a time when factorial studies of intelligence and mental test score performances were beginning to challenge the more generally held contention of "g" or a central component of overall intelligence. The authors of the California Test of Mental Maturity acknowledged their theoretical indebtedness to the findings of the multiple factor analysts. They also seemed to accept the principle of specificity and independence of differing kinds of intellectual behavior such as memory, spatial relations, reasoning, number facility, and verbal ability. In their descriptions of the tests the authors recognized the need for differential analysis of scores on tests, but showed sharp differences in the combination of "factors" that went to make up the Language and Nonlanguage IQ's and MA's that this test purports to measure.

The need for diagnosis of intraindividual trait differences was called to the reader's attention in 1931 by the writings of one of the authors of the California Test of Mental Maturity.¹¹ He indicated that the problem of measuring trait differences was becoming increasingly important. In a revision of this work, he further stated: ". . . to be useful in the diagnosis and solution of learn-

¹⁰ *Manual for the California Test of Mental Maturity*. Los Angeles: The California Test Bureau, 1937.

¹¹ Ernest W. Tieg, *Tests and Measurements for Teachers*. Boston: Houghton Mifflin Co., 1931, pp. 44 and 295-296.

ing difficulties, intelligence test data must reveal the ways in which different students learn and the reasons for their failure to learn effectively. . . . Research has done much to clarify this situation and reveal that such factors as perceptual speed, memory, comprehension, mathematical ability, inductive and deductive reasoning, and verbal abilities are combined in various ways to produce different as well as identical IQ's . . ." ¹²

The author then acknowledged the factor analysis work of such men as Hotelling, Kelley, Thurstone, and Guilford in a footnote.¹³ At the same time he stated that the California Test of Mental Maturity was developed to bring out more or less independent yet interrelated abilities, so that a teacher might analyze and evaluate each group of *factors* of mental maturity separately. In the 1947 edition of the manual a statement was made that the authors ". . . believe that the multiple *factor* theory of intelligence comes nearer to explaining observable phenomena than does the strong central-factor theory alone . . ." In the 1950 edition of the manual and subsequent revisions the reader is told that the theoretical framework for the development of the CTMM was based upon Elizabeth T. Sullivan's "Psychographic Record Blank."¹⁴ Her instrument was designed to analyze Stanford-Binet performance and to set up some 14 categories into which response items might be grouped. It was from this "conceptual framework" that items were developed for what was to be named the California Test of Mental Maturity. Further statements were made in the 1951 manual that test data from these items were then ". . . factor analyzed by the Thurstone Centroid Method. . . ."

The selection of items found in the California Test of Mental Maturity was a subject of further comment in a statement by one

¹² Ernest W. Tieg, *Tests and Measurements in the Improvement of Learning*. Boston: Houghton Mifflin Co., 1939, pp. 39.

¹³ An almost identical reference to the same sources on factor analysis is made in footnotes in the several editions of the California Test of Mental Maturity manuals, 1936 to 1950.

¹⁴ California Test Bureau, *Psychographic Record Blank*. Los Angeles: California Test Bureau, 1926.

of the authors who said that its items ". . . *obviously* sample a wide variety of relationships involving immediate and delayed recall, spatial relations, logical and numerical reasoning, and verbal concepts . . ." ¹⁵

Careful examination of each of the subjects indicates to the counselor that his counselee must do the following tasks to achieve scores in the subtests named.

TABLE 15. Tasks Set by Subtests

Name of Subtest	Task
Test 1. Immediate Recall	Recall the second of a series of words after original oral reading in couples.
Test 2. Delayed Recall	Select from multiple choice items details of a story read orally 30 minutes before.
Test 3. Sensing Right and Left	Identification of lefts and rights from among twenty pictures of hands, feet, ears, gloves, and wings in differing positions.
Test 4. Manipulation of Areas	Identify abstract spatial patterns similar to paper form boards.
Test 5. Opposites	Identify odd and opposite pictures among pictorial items of similarities and one opposite.
Test 6. Similarities	Identify correct and similar pictorial items from among several pictures.
Test 7. Analogies	Choose among pictorially presented analogies.
Test 8. Inference	Select from verbal syllogisms with multiple-choice options.
Test 9. Number Series	Identify incorrect number from patterned serial presentation of numbers.
Test 10. Numerical Quantity	Determine correct coins needed to make predetermined amount.
Test 11. Numerical Quantity	Solve word problems in arithmetic with multiple-choice options.
Test 12. Verbal Concepts	Identify correct vocabulary items with multiple-choice options.

¹⁵ Ernest W. Tiegs, "The Proper Use of Intelligence Tests," *Educational Bulletin*, No. 14, Los Angeles: California Test Bureau, 1945, 1951.

In later revisions of the manual of the California Test of Mental Maturity the authors state that the tests were designed to measure more of the types of intellectual processes than the Stanford-Binet was designed to measure. The evidence they offer to support this theoretical contention is a short statement that the total scores of the California correlates as high as .88 with the Stanford-Binet.

It should be obvious from the illustration given above that the basic assumptions made by the authors of a test will determine the form of the test. The counselor must examine each set of assumptions very carefully and decide which one is most acceptable to him and which is most likely to provide results that can be interpreted meaningfully to his counselee, his parents, his current or prospective teachers, and potential employers.

As he considers IQ scores on a student's record the counselor may decide that though the scores are derived from different assumptions and processes, he can compensate for the resulting difference. Most mental ability tests suggest that scores can be used to assist pupils to choose wisely in planning their educational and vocational programs, and many counselors may accept such statements uncritically. If he accepts the results of the tests at their face value, he may find himself saying of a low score that the student just "doesn't have it, and there is not too much use in pressing him," or, "I must try to dissuade this youngster from his aspirations," or "College is obviously beyond this student." Or, of the counselee who makes a high test score but achieves low marks the counselor might say, "Obviously, this student is falling down on the job. He can do much better." The consequences of such interpretation can be serious in terms of a youth's future. From the notation "IQ, 100" may come a decision that action in the direction of motivating the student to do better would be fruitless, or a decision to try to get the student to change his program from college preparatory to something "easier." Or perhaps he may consider the desirability of recommending that the student be placed in a "slow" section of some class where he will not be expected

to show much progress. Before he makes such decisions, however, the counselor may want to check his "score" more closely and continue consideration of factors that may influence it.

VALIDITY REPORTS IN TEST MANUALS

No one single criterion for the evaluation of a psychological test is as important as validity. In the final analysis factual data, supported by research studies that show how well the given instrument does what it purports to do, are of much greater importance than all the other commonly used criteria combined. In counseling with individual students, *predictive validity* is the type of validity data that seems to be most useful. Granting the importance of such considerations as reliability, ease of scoring, readability, costs, time for administration, adequacy of norms, and care in standardization, all these come to naught unless specific and exact information about predictive validity is presented in the test manual. It should give assurance to the counselor who proposes to use the test that the traits, "factors," or skills he thinks he is measuring have some relationship to identifiable criteria that have meaning for the case in hand.

If it was well established that a test did what it was supposed to do, the points noted above need not concern the counselor. Test manuals usually suggest that mental ability tests "can be used effectively" for the following purposes:

1. Classification of pupils according to their *ability* to profit from education.
2. Guidance of *individual* pupils in making educational and vocational choices.
3. Adapting levels of instruction to the *ability levels* of individuals or classes.
4. Interpreting levels of achievement of *individuals* or groups.
5. Providing research information about the mental levels of groups or *individuals*.

6. Obtaining *diagnostic* profiles to aid teachers and counselors in discovering learning difficulties.
7. Obtaining clues for educational and vocational *guidance* from separate factor scores.
8. *Predicting* success in certain fields of work.
9. Providing evidence of *mental deterioration*.

Such lists as the above imply that mental ability tests can do much to assist a counselor in his work with individual students and in his attempts to assist other members of a school faculty in reaching common goals. But the counselor cannot accept such statements unless they are supported by empirical evidence. Unless it is available he must resist the temptation to predict his counselee's chances to deal successfully with global concepts, rely on textbooks, be successful in recipe construction, or do well in drama—all specifically mentioned in commonly used test manuals. He must inhibit tendencies on his part to spot diagnostic signs of mental deterioration, lack of preciseness in relations with others, or difficulty with music on the basis of performance on a test. He might look in vain for even a simple study that shows the relationship of scores on tests and subsequent teachers' marks—the usual criteria offered as proof of predictive validity for nearly any such test.

It is difficult to believe that a major test publisher or well-trained test builder could market a test that offers absolutely no empirical evidence that it does the job its authors claim for it, but unfortunately no test manual contains evidence that is fully convincing on any of its claims. Some test authors admit that the validity of any mental test is difficult to establish and, providing no data, indicate that evidence of validity must wait for further knowledge about mental development. Others present general statements to the effect that the test has been used successfully over a long period of time and indicate that in *some* situations students who have graduated with honors made scores in the highest ranges of the test. Still others suggest that a mental ability test is valid if it indicates the probable rate of progress a pupil will make in getting through

school. Many authors beg the question and give as chief evidence of validity that the test scores correlate highly with another that bears a similar title.

The use of a patent-medicine-advertising form for presenting evidence of validity must come as a rude shock to anyone who has been led to believe, as a result of the statistical data presented in the section on the construction of the test, that he was dealing with a scientific instrument. Because a test has been successfully used (no data given to amplify the word "successful") the counselor is supposed to accept the statement that it is good. Does usage always assure value? Although it is claimed that the test is effective in doing the nine things described above, there is only evidence that in *some* situations honor graduates made scores in the highest range of the test. How many are *some* situations? Does the use of *some* imply that, in others, the conditions did not hold? What about progress through school as a validity criterion? What factors other than brightness may influence it? Could such factors as health, frequent change of school and the attendant problems of adjustment and variability of curriculum, be important in determining progress?¹⁶

If a counselor were to look for straightforward evidence that mental ability tests can do what their authors suggest, he may be somewhat upset as he is led through series of confusing rationalizations that ultimately prove to be less than satisfying. He must reject many of the tests for his own protection and for the protection of his counselees. Consider the difficulties he might get into if he were to try to interpret some test scores to a bright senior high school student and his parents with whom enough rapport has been developed so that they felt free to ask penetrating questions. How would he answer them when they asked for evi-

¹⁶ Durrell's suggestion that "at least 25 percent of the children who make slow progress in school are of normal or superior intelligence" may cause the counselor to further question validity evidence in the form of rate of progress in school. Donald D. Durrell, "Learning Difficulties Among Children of Normal Intelligence." *The Elementary School Journal*, December, 1954, 55:201-208.

dence about how well the test did what it was supposed to do? His answer that it had been used "successfully" would elicit a further questioning on the definition of "successfully," which could be answered only by the vague statement that in *some* situations honor graduates made highest scores. And he would not have the slightest evidence about those situations or where they might be found. He would have no evidence with which to answer their questions about the value of tests in classifying students, or guiding them in making educational and vocational choices that are consistent with their educational level. The poker-playing parent would probably conclude that he had caught the counselor bluffing.

The problem of variability of performances in the various tests, too, may cause considerable difficulty in interpretation. The selection of a test for measurement of "mental ability" can mean much when a counselor is deciding what he has and what he can do with it.¹⁷ Some of the variability in yield of tests was indicated earlier and Travers writes to the same point: "Intelligence quotients from different group tests may measure somewhat different aspects of intelligence . . . variations in the psychological process involved, and the varying emphasis on numerical, perceptual, and verbal materials, may seriously affect the uses to which the scores might be put."¹⁸ One example may illustrate the point as it applies to mental ability tests. If a counselor examines one test he may find that he has selected a "verbal test of mental ability" that includes, among others, items that require arithmetic reasoning. If the counselor had chosen instead the Terman-McNemar Test of Mental Ability, he would also have had a measure of "general verbal intelligence," according to the manual for that test, but in the latter instance the arithmetic and numerical items were expressly omitted. That the choice among available tests of mental ability may well influence the score or IQ that is recorded on the

¹⁷ Roger T. Lennon, "A Comparison of the Results of Three Intelligence Tests." Test Service Notebook, No. 11, Yonkers, N.Y.: World Book Co.

¹⁸ Robert M. W. Travers, *Educational Measurement*, New York: The Macmillan Co., 1955, p. 347.

cumulative record should now be clear to the counselor. Though the designated measure, *mental ability*, would still be indicated by a particular test title, the actual results and implications for counseling might vary greatly.

The counselor cannot avoid this confusion and must wonder at times what he has when a test score has been obtained. Of course he may feel that it might not matter greatly what differences he finds between tests if the one he has selected for use does what the manual says it does and what he wants it to do.

In view of the fact that counselors would be well advised to avoid the use of tests for which there is no empirical evidence of validity presented in the form of expectancy tables, it would not be necessary to consider additional facts about them. Since, however, he may find satisfactory evidence of validity about a test he should then go on to further study of its characteristics. He may continue to look at his test to appraise additional factors such as those described below that must be considered in choosing a test and interpreting the scores he obtains from it.

TRYOUT AND NORMATIVE PROCEDURES

Since a score that the counselor is considering represents a comparison of his counselee's performance with that of others in similar circumstances, he must make a thorough investigation of the norms reported in any test manual. Being somewhat acquainted with the concept of norms, he may wonder, for instance, whether his counselee can be compared to the group represented by the norm. He will want to know about the communities and their representativeness. Did the norm group represent a chance and convenient population or a random selection? Are there any sex differences?

Much of the norm data presented in test manuals are scanty and in many cases the wording in the section on norms is confusing. In the manual for one test, for example, it is stated that the norms

for this test were established through a national testing program. Approximately 190,000 tests were *distributed* to 200 communities in 37 states and 307 parochial schools in the diocese of a certain city. In a footnote, however, it is noted that: ". . . not all communities returned their tests in time to be included in the normative population. The norms are *based on* the results from 148 communities in 33 states where answers were recorded in test booklets." (*Italics added.*)

Just why the authors indicate in the main body of the text that 190,000 tests were distributed to 200 communities in 37 states is difficult to understand. What happened to the results from the 307 parochial schools? Are they incorporated in the normative population, and does it make any difference to the counselor whether or not they were used? He is not told the states in which the 148 cooperating communities may be found although it is indicated later that there was a "wide geographical distribution."

Such descriptions of norm groups must leave the counselor with many doubts. He will want to know if the participating communities were rural or urban or mixed? Were the pupils enrolled in all the usual kinds of curricular programs? What is the relative distribution of boys and girls, especially in the upper grade levels where there may have been more dropouts? What was the socioeconomic status of the communities? And what happened to those students in the parochial schools who received special attention? It is possible that the test authors did obtain a good cross section of high school youth, but the evidence that they did so is lacking.

After the meager and inadequate description of the populations the authors go on to point out that, in order to facilitate calculations, actual computations were based on only a 10 percent random sample of the group tested. They then indicated that "the *wide geographical distribution* of the cooperating communities and the fact that all students in at least three consecutive grades were tested in each community *should* insure a cross section of the school

population in the grades involved in the norms." (*Italics added.*) They *should* but *did* they?

Again the counselor must ask about the adequacy of a 10 percent sample. Was it enough? Was it truly random, since the method of getting randomness is not described? If random numbers tables were used, did the procedure really get proper samples of ethnic, sex, and socioeconomic groups? Does the taking of one sample by random numbers methods really assure randomness? Were the risks involved in the procedure of taking a 10 percent random sample of an inadequately described population worth taking merely to facilitate calculations? What will the counselor say to a student or his parents if they question the results and have to be told that the counselor himself questions them because, in order to facilitate calculations, the test authors used questionable procedures?

Some publishers of tests simply indicate the size of their norm groups without presentation of enough evidence of the characteristics of the subjects to make meaningful interpretation possible. Nowhere in the literature issued by the publisher of one widely used test is there a clear-cut description of the original normative sample used in standardization nor is there any description of this group's important characteristics, its location, socioeconomic composition, or selection. Counselors working with individuals in various sections of the country and looking at these norms might raise questions as to how appropriate they are for their particular counselees. The counselor might be a little wary of using norm tables, too, without some information about such factors as these: presence or absence of rural or urban youth, selection by type of training institutions in the normative group; socioeconomic composition of the sample; and the achievement level of the groups included in the sample. Pertinent information of this sort is too often conspicuous by its absence. Nor is there any explanation as to how items, for which no normative data are furnished, suddenly appear in diagnostic profiles.

Until more adequate normative information is provided by test-makers the counselor must often rely upon his *interpretations* of the meaning of his counselee's scores. He may have to augment this scanty norm data by his own follow-up of students and local normative information as long as meaningful norms are lacking.

RELIABILITY REPORTS IN TEST MANUALS

The authors of many tests indicate that "the reliability of a test is the stability of the measures it yields." The counselor must take particular notice that the word "reliability" is used in its technical sense, defined as above, rather than in the usual dictionary meaning of the word as "the state of being reliable," when the word "reliable" is defined as "trustworthy." It is conceivable that a test may be highly reliable in the sense that it yields stable measures (the length of period of stability is not given) while at the same time it may not be trustworthy in counseling for such reasons as those given above in the section on validity. When the counselor or counselee plans to take some action such as hypothesizing about the future or helping to plan and make decisions, he is confronted with the question as to how closely current performance on a test correlates with later performance. He is rarely interested in momentary measurement, but he is concerned about consistency of performance over a long period of time. Test authors rarely if ever provide such evidence of long-term consistency. They rely rather on evidence of consistency in performance at one sitting or on two that are completed within very short intervals.

One test manual reports split-half coefficients for a population of nearly 300 students in Grades 7 to 9. Interform reliability was computed for less than 250 students *in the same grades*, and the authors stated that coefficients for other grade ranges (unspecified) varied only slightly from the 14-year-old age range. The counselor would note in this case that other age ranges are not specified and that they varied an unspecified amount from the

14-year value. Having all these statements, the counselor has three coefficients computed for lower age and grade levels with subjects about whom he knows next to nothing except that they lived in a certain state. He will wish that the authors had given him some more complete tables and further information on their subjects.

The coefficients described above are only vaguely and generally helpful to the counselor when a counselee asks if he may take the test over again because he feels sure that he could do better next time. He might, since coefficients cannot be interpreted to most counselees and their parents, try to get some help from the report on the standard error of measurement. This in a certain test is reported as "approximately 3.2 standard score points for the entire range covered by the test." The counselor may be tempted to answer the counselee's question by indicating that, in general, on the whole, on the average, and other things being equal, if he is somewhat like seventh to ninth grade pupils in certain cities in a certain state it is not likely that a second trial of the test (at some unspecified time) will yield scores that differ greatly from those that the first trial yields. Such necessary vagueness will not be comforting to the eager inquiring counselee.

It is common practice to use only a small sample of an original population in the computation of reliability coefficients. In one test, for example, less than 1 percent of the subjects in the norm group were used. No description of the sampling process is given, so the counselor might wonder just how the subjects were selected from a parent population more than 100 times as large. He might legitimately wish to have information as to their geographic location, type of schooling, sex, socioeconomic, racial, or other characteristics, and he might wonder what sort of sampling procedures were used in their selection. And although the variability of their scores in this test is reported in MA units ($S.D. = 23$ to 32) no information is given as to their average scores, their level of performances, or their educational achievement. In the same test no reliability coefficients were offered for the various subtests of the

test, although users were urged to use a diagnostic profile of scores. Comparison of scores or subscores whose reliability is not known cannot be a useful procedure in counseling.

It is common practice, too, to present reliability coefficients for several grade levels combined. This type of sampling which allows for maximum variability tends to produce a much higher reliability coefficient than one that is determined on a single grade level.¹⁹ In this approach the counselor is left with the untested assumption that the reliability estimate for his individual counselee who is in, say, the eleventh grade, will be similar to the data reported for a much more variable group in Grades 9 to 12. He is left with the responsibility of determining the real reliabilities for his own local population.

The good counselor will not be naïve in his interpretation of the standard errors of measurement that often appear in the reliability sections of test manuals. In one manual, for example, one may find the statement that the standard error of measurement was 3 points and that a pupil's score will be in error not more than 3 points 662/3 percent of the time. This statement may seem very good, and the counselor might interpret the information to mean that, for his counselee whose IQ score was 100, the chances are two to one that on successive retests the IQ score would fall somewhere between 97 and 103. This rather narrow range would seem to suggest that the obtained IQ score was rather stable. He must remember, however, that this figure applied only in general and may not apply in the case of his particular counselee. He must also consider the possibility that there is still one chance in three in the general situation that the IQ would go beyond that range. He would have to keep in mind, too, that the reported standard error of measurement covered a range of five grades. Since it would usually be a higher error figure for a single grade, the counselor might well wonder about the size of the standard error of measurement for the single grade in which his counselee happened to be.

¹⁹ Anastasi, *op. cit.*, pp. 115-117.

It is possible that the factor of reliability in some measurement may be overemphasized. Where the interpretation of the results may have long-range implications, the stability of the score becomes important.²⁰ That would appear to be the case with many tests since one of the assumptions underlying them is that ability to learn is a fairly constant quality. Some authors have placed major importance on the assumption of "constancy" when they write that ". . . degrees of 'brightness' are theoretically constant for a given child, being a fixed characteristic of the endowment, so that the child who is really below normal at one age will be so at all ages, and that an adult who is above normal was so much above normal, relatively, at any age of development. . . . Its truth is, in fact, the foundation of our hopes in testing intelligence, for if it is not true, in part at least, we cannot prognosticate, and intelligence measurement will be of no great value."²¹

Written some 40 years ago, this concept is still a very important hypothesis underlying many tests. It is a fundamental assumption of those who use tests in attempts at prediction. This being the case, adequate evidence of the stability of scores is of more than passing concern to the counselor.²²

²⁰ Henry E. Garrett, "A Developmental Theory of Intelligence," *American Psychologist*, September, 1946, 1:372-78.

²¹ Arthur S. Otis, "A Criticism of the Yerkes-Bridges Point Scale, with Alternative Suggestions," *Journal of Educational Psychology*, March, 1917, 8:129-150.

²² It may be of interest to summarize some research done on this point and reported by Traxler. (Arthur E. Traxler, "Reliability, Constancy, and Validity of the Otis IQ," *Journal of Applied Psychology*, April, 1934, 18 241-251.) Traxler reported on the Otis Self-Administering Tests of Mental Ability, Higher Examination. The reference is not inappropriate, since subsequent forms, the Gamma Em included, appear to be built upon this earlier test. Traxler noted that "Unless the IQ secured from a group test is highly dependable, marked injustice may be done in the classification and grouping of individuals because of the false information relative to mental ability." Commenting further on the significance of this point, Traxler stated that ". . . in view of the fact that IQ's found for a class at the time of entrance to high school are frequently recorded and used for several years by the school, the correlation between forms of a test administered in successive years is a matter of considerable importance." With small groups ($N=85, 100, 75$) graduating from the University of Chicago High School, Traxler found stability coefficients ranging from .647 to .807, the mean of ten coefficients being .725 between forms administered a year or two years apart. Regarding the constancy of the IQ, Traxler reported that, of 885 changes studied, one fourth were of nine points or more and

DIRECTIONS FOR ADMINISTRATION OF TESTS

Authors of tests usually present to the persons who are to administer tests admonitions concerning avoidance of influences that might cause tenseness and anxiety, the guarding against interruptions, the provision of supplies, and the need to become familiar with the instructions. It is often suggested that the time limits be adhered to within a margin of a *few* seconds, but *few* is frequently not defined. There is occasionally a suggestion that time limits should not create a feeling of pressure or nervousness on the part of the pupil, but those who have administered such tests know that close timing of tests is likely to create the tenseness anxiety supposed to be carefully avoided. And despite the cautions about timing, some authors state that it was their intention to make a power test rather than a speed test. In some cases it is suggested that in above-average groups an entire class will finish a subtest before the time for it has expired. Instead of using this time to let the subjects go back to check their work, some directions suggest that the examiner should continue with the directions for the next subtest. It is possible for one student under such circumstances to influence the whole testing procedure since, if he is the one who has not finished, the examiner must wait (up to the limit of the "time—within a few seconds") until all have finished. Some discretion is often left to the examiner and when that is permitted an opportunity for error in test administration is possible. There is also some possibility for error when separate answer sheets are used, since after some tests are completed the students are asked to retain them to blacken their marks and erase any stray ones.

that one was as high as 25 points. Though noting that, when compared with constancy data of other group tests, the Otis change was relatively small, he stated that "... about one Otis IQ in six changes so rapidly that marked inaccuracies might occur if one test alone was used for purposes of classification or grouping in high school. Even more marked individual IQ score variability was reported by Nancy Bayley. "Consistency and Variability in the Growth of Intelligence from Birth to Eighteen Years." *Journal of Genetic Psychology*, December, 1949, 75:165-196.

Discussions of the effect of such factors as fatigue, health, test-wiseness, and distractions were presented in detail in Chapter III. It will suffice here to remind the reader that such factors may influence scores whenever tests are given, and that some of them may be a function of the way the test is administered. They will not be evident from the usual entry of test scores or a cumulative record.

It is possible, of course, that the student was in "top condition," that the "human factors" were all favorable, but that all was not well in the *administration*. The counselor can only speculate as to which of many influences upon a test score may have affected his counsellee's score if the directions were not followed precisely. If he is especially sensitive to such matters as these, however, he may be a little disturbed when he reads suggestions in some manuals that the tests are self-administering and that it is merely necessary to pass out the booklets, allow the pupils time to study the first page with a minimum of directions, and let them go ahead and take the test.

The casualness of such an approach may appear to the counselor to be more than a little perfunctory and he may wonder about the outcome of "a minimum of directions." There may be definite advantages, in the press of other work, to test all the pupils in a school in a day, as is proposed in one manual, but it may occur to the counselor that the proposed approach, rather than assuring "reasonable uniformity," may actually encourage considerable variability unless the teachers left in charge are as well trained as he is.

As he inspects the directions that the students are to read he may wonder further what his students' reaction to the following might have been. "This test contains 90 questions. Do the best you can, though you are not expected to be able to answer all of them. After the examiner tells you to start, you will be given a half hour. Answer as many questions right as possible. Do not go so fast that you make mistakes. Do not spend too much time on any one question. No questions about the test will be answered by the examiner after the test begins."

Did the student heed all this advice? Since he was not supposed to be able to answer all the questions, did he relax and take it easy? Or, since the task looked somewhat staggering for 30 minutes, did he tighten up because of his apprehension? And since not all the types of items found in the test are demonstrated in the directions, can the counselor assume that he understood what he was to do on each new one as he came to it? The examiner's oral directions call for asking, "Is there anyone who does not understand how to answer the samples?" But was the student one who, in a large group, did not want to let others know that he did not understand what he was to do? If the answer to some of these queries is positive, the counselor may have revealed more factors that can influence test scores.

DIRECTIONS FOR SCORING AND RECORDING OF TEST SCORES

As the counselor looks at a test result on a cumulative record, he may recognize the importance of accuracy in transcribing the score and IQ from the test answer sheet. He may be inclined to assume that it was done correctly, but errors have been known to occur. The implications of the misreading of a number or the inadvertent recording of a score on the wrong cumulative record need no further comment. The recording of the score is preceded by the computation of the IQ and, in the case of some tests, this is relatively simple procedure.

But assuming that the recording and computation of the score have been done accurately and checked well, the counselor may continue in his reverse chronology of the development of a test score and speculate on the possibilities of an error in the initial scoring of the answer sheet. Hand scoring can become tedious if done for any period of time, and an error may not be recognized in the deceptively innocuous-looking "number" on the cumulative record form.

One of the claimed features of many recently developed tests

is the rapid method of stencil-scoring. A test is scored by placing a single punched stencil over the answer sheet so that only the right answers appear. The total score is the number of marks appearing through the punched holes. Such an arrangement assumes that the students have followed the directions, "Never put more than one mark in any row of spaces." This, of course, leaves something to be desired, for it is *possible* that an enterprising student, *unable to reduce the alternatives to less than two, may check both*. The right one may show through the punched hole while the wrong answer remains covered. Some authors have anticipated this possibility by suggesting in the manual that "if in the case of any item two marks have been put in the same row of spaces, draw a colored line through the row of answer spaces and allow no credit for that item." The recognition of the above possibility goes only part way in that it suggests that this be done if one *happens* to notice several marks. The directions should be more explicit and should require the test users to screen every answer sheet before scoring with the stencil. To make such a positive statement, however, would result in an extension of scoring time and a reduction in one of the suggested major features of objective tests—quick-scoring. Another approach, of course, would be to provide a "wrongs" stencil through which all but the right answers would appear. This, too, would require another handling of answer sheets and again eliminate the quick-scoring feature to some degree. If the test administrator were conscientious and were aware of the above possibilities, it seems likely that he would sacrifice speed for accuracy, *though not specifically told to do so*. If not, scoring errors of this kind might conceivably be represented in the score on the student's record.

Provisions are made for machine scoring of many tests and it is assumed that anyone who attempts machine scoring will be thoroughly acquainted with the technique, which may be reasonable enough. The following quotation from one test manual suggests

to the counselor some of the possibilities for scoring error if a machine is used.

Scan *each* answer sheet carefully before it is scored. [This would have been a sound recommendation for hand scoring too.] Where more than one answer has been marked for an item erase all marks for the item. Erase any stray marks made in the answer space, inasmuch as even very small and light marks are sometimes sensed by the machine. If the pupil has failed to make complete erasures, make a clean erasure. If the marks are too light, go over them with one of the special lead pencils. Check carefully by hand a certain proportion of the answer sheets to insure maximum accuracy.²³

With all these possibilities for error in machine scoring, the counselor will not want to assume too much regarding the accuracy of his test score unless he has taken all these precautions or is confident that they were taken by the person who administered the test.

Going back another step in the process and assuming that the score has been properly transcribed and accurately scored in the first place, the counselor may well consider the possibilities of the effect of the *human factor* in the score he is contemplating. Here such things as the attitude of the student at the time he took the test, his health, comprehension of directions, and other factors can conceivably be represented in the "number" without being evident unless noted specifically on the record.

As the counselor looks again at test manuals he will consider the assumptions and procedures involved in scoring. He will find that many test forms are designed to permit more rapid scoring by a perforated key and he must wonder if the gains in speed of scoring have resulted in obtaining less valuable scores more rapidly. No clear-cut evidence on this point is available. The counselor may wonder, too, regarding the scoring of a test, by what feats of insight a test-maker can decide that each item in the test has exactly the same value as another. Studies in the value of weighting items

²³ *Manual for the Otis Quick-Scoring Mental Ability Test*, op. cit., p. 5.

are so inconclusive that the allotting of equal values to all items must still be considered a questionable procedure. If, of course, the validity of a test is high there will be no need to question the scoring procedure, but since, in many tests, the validity is questionable, the scoring procedure must be one of the factors that is suspect.

SUMMARY

In this chapter the factors inherent in a test score, factors which may influence the score, and the implications of the factors for the interpretation of scores have been presented. It has been pointed out that the *counselor needs to consider these factors both in the initial selection of a test and, ultimately, in the appraisal of the scores he obtains from the tests he has selected*. It has been noted that each test available to the counselor is different, and that the value and usability of the test results will vary accordingly. In effect, the discussion in this chapter illustrated practical applications of the criteria for the selection of a test presented in the previous chapter. The examination processes described in this and the following chapter are those that counselors, in coöperation with other members of a school staff, might employ in the selection of a test for use in counseling. It appears likely that thorough study of a test in the manner described here would result in recognition of the limitations and values of tests in helping counselees to help themselves to make important personal decisions.

DISCUSSION QUESTIONS AND EXERCISES

1. Select three mental ability tests and examine the manuals for concepts of mental ability represented. What differences or similarities do you find in the concepts? Do the *differences in concepts* appear to be reflected in the types of items used in the respective tests? Do the test authors present evidence in support of their particular concepts of mental ability?
2. It was suggested in this chapter that the group upon which a test

author tried out his test items would have bearing on ultimate test scores earned by individual students. Why is this the case? What differences would you expect to find in such item trials between groups selected from traditional Eastern college preparatory schools and groups selected from small, bilingual communities of the Southwest? Between rural and urban groups?

3. The manuals of some tests of mental ability suggest that the test results may be used to assist students in the choice of an occupation. What evidence would you wish to have in substantiation of such a claim? Do you think occupations can be classified in terms of the mental ability required? Why or why not?
4. Prepare a critical review of a test of *mental ability* based on the pattern presented in this chapter. After completing the review, compare your review and analysis with that of reviewers in the *Mental Measurement Yearbooks*. To what extent are you in agreement? Can you defend your stand on points of disagreement? Are the reviews in the *Yearbook* consistent with each other? What factors might account for differences among the *Yearbook* reviews?
5. It has been said that "the old tests are the best." On what assumptions might such a statement be based? Do you agree with the statement? How might the "age" of a test influence the test scores of today's youth?
6. Some attempts have been made to set up scales for the evaluation of tests on 100-point scales. In them 25 points may be given for validity, 20 points for reliability, ten for clarity of instructions, etc. Is any such system defensible? Why or why not?
7. Select from your test library any widely used *aptitude* test other than one that has been considered in this chapter. Using the list of factors noted in this chapter, write your own evaluation of the test. When you have finished compare your report with the reviews of it in *Buros' Mental Measurement Yearbooks*. What factors may have produced differences between your evaluation and those of the reviewers in the yearbooks?
8. Choose five tests that are recommended for use in counseling. If there is a statement in their manuals that scores may be used in

educational and vocational guidance, examine and report on the evidence offered to justify the statement.

REFERENCES

- American Educational Research Association. "Psychological Tests and Their Uses." *Review of Educational Research*, February, 1947, 17:1-128.
- American Educational Research Association. "Educational and Psychological Testing." *Review of Educational Research*, February, 1953, 23.
- American Educational Research Association. "Educational and Psychological Testing." *Review of Educational Research*, February, 1956, 26.
- Anastasi, Anne. *Psychological Testing*. New York: Macmillan, 1954.
- Buros, O. K. *The Fourth Mental Measurements Yearbook*. Highland Park, N.J.: Gryphon Press, 1953.
- Coleman, William, and Cureton, Edward E. "Intelligence and Achievement: The 'Jangle Fallacy' Again." *Educational and Psychological Measurement*, Summer, 1954, 14:347-351.
- Cronbach, Lee J. *Essentials of Psychological Testing*. New York: Harper, 1949.
- Crowder, Norman A. "The Holzinger-Crowder Uni-Factor Tests." *Personnel and Guidance Journal*, January, 1957, 35:281-287.
- Doppelt, Jerome E. "Progress in the Measurement of Mental Abilities." *Educational and Psychological Measurement*. Summer, 1954, 14:261-264.
- Dreger, Ralph M. "Different IQ's for the Same Individual Associated with Different Intelligence Tests." *Science*, December, 1953, 118:594-595.
- Durrell, Donald D. "Learning Difficulties Among Children of Normal Intelligence." *Elementary School Journal*, December, 1954, 55:201-208.
- Freeman, Frank S. *Theory and Practice of Psychological Testing*. (Revised Edition.) New York: Holt, 1955.
- Garrett, Henry E. "A Developmental Theory of Intelligence." *American Psychologist*, September, 1946, 1:372-378.

- Kelley, Truman L. *Interpretation of Educational Measurements*. Yonkers, N.Y.: World Book, 1927.
- Knezevich, Stephen. "The Constancy of the IQ of Secondary School Pupils." *Journal of Educational Research*, March, 1946, 39:506-516.
- Lennon, Roger T. "A Comparison of Results of Three Intelligence Tests." *Test Service Notebook*, No. 11. Yonkers, N.Y.: Division of Test Research and Service, World Book (Undated.)
- Lindquist, E. F. (Ed.). *Educational Measurement*. Washington, D.C.: American Council on Education, 1951.
- Otis, Arthur S. "An Absolute Point Scale for the Group Measurement of Intelligence." *Journal of Educational Psychology*, May, 1918, 9:239-261.
- Rulon, P. J. "On Concepts of Growth and Ability." *Harvard Educational Review*, Winter, 1947, 17:1-9.
- Schmidt, Louis G., and Rothney, John W. M. "Relationships Between Primary Mental Abilities Scores and Occupational Choices." *Journal of Educational Research*, April, 1954, 47:637-640.
- Stewart, Naomi. "AGCT Scores of Army Personnel Grouped by Occupation." *Occupations*, October, 1947, 26:5-41.
- Super, Donald E. *Appraising Vocational Fitness*. New York: Harper, 1949.
- Super, Donald E. "The Tests of Primary Mental Abilities. Comments." *Personnel and Guidance Journal*, May, 1956, 35:577-578.
- Super, Donald E. "The Holzinger-Crowder Uni-Factor Tests. Comments." *Personnel and Guidance Journal*, January, 1957, 35:287-288.
- Thorndike, Robert L., and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education*. New York: Wiley, 1955.
- Thurstone, Thelma G. "The Tests of Primary Mental Abilities." *Personnel and Guidance Journal*, May, 1956, 35:569-577.
- Travers, Robert M. W. *Educational Measurement*. New York: Macmillan, 1955.
- Traxler, Arthur E. *Techniques of Guidance*. (Revised edition.) New York: Harper, 1957.
- Traxler, Arthur E. "Reliability, Constancy, and Validity of the Otis IQ." *Journal of Applied Psychology*, 1934, 18:241-251.

CHAPTER V

The Use of Standards in Test Selection

Reference has been made earlier to the publication in 1954 of the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*.¹ It was pointed out in Chapter III that the publication of these recommendations represented an important event in the area of measurement. The recommendations suggested that many tests had serious shortcomings and pointed to the way in which improvements might be made.

The need for standards of some kind was suggested by the authors of the recommendations in this statement.

Professional workers agree that test manuals and associated aids to test usage should be made complete, comprehensible, and unambiguous, and for this reason there have always been informal "test standards." Publishers and authors of tests have adapted standards for themselves, and standards have been stated in textbooks and other publications. . . . Until this time, however, there has been no statement representing a consensus as to what information is most helpful to a test consumer. In the absence of such a guide, it is inevitable that some tests appear with less adequate supporting information than

¹ *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Supplement to the *Psychological Bulletin*, March, 1954, 51:1-38.

others of the same type, and that facts about a test which some users regard as indispensable have not been reported because they seemed relatively unimportant to the test producer. This report is the outcome of an attempt to survey the possible types of information that test producers might make available, to weigh the importance of these, and to make recommendations regarding test preparations and publication.²

Many tests appeared before the thinking of theorists and professional groups crystallized into a definite statement of standards for psychological tests. The manuals accompanying these tests presumably represented what the authors and publishers regarded as adequate information about the tests and their use. They could not have been expected to anticipate all the recommendations that appeared at a much later date because they represent the thinking of many persons and are the product of an assembly of measurement "talent" not available to individual publishers. Since they are now available, publishers may well review all test manuals published before the Technical Recommendations appeared with a view to revision in the light of the new standards.³ In any case it seems reasonable to expect that test manuals published after the recommendations became available should reflect the standards presented therein. The degree to which this expectation has been met in the case of one such test and manual will be discussed in the following pages by making a direct comparison of statements taken from the test manual and from the standards in the Technical Recommendations report.

Before the test manual is examined in the light of the Recommendations a discussion of their development, scope, general organization, and content will be presented. It may aid the reader in

² *Id.*, p. 1.

³ This is not to suggest that standards were not available at the time of publication of these tests. The materials included in the Technical Recommendations do not reflect recently developed concepts. They represent a consolidation of much that has been known for many years.

following the analysis of a test and in gaining a better appreciation of the importance of the document itself.

TECHNICAL RECOMMENDATIONS FOR PSYCHOLOGICAL TESTS AND DIAGNOSTIC TECHNIQUES

ORIGIN

The recommendations were designed to provide guides for the test author in development of tests, for the publisher in preparing and presenting data needed for effective use of tests, and for the person who must ultimately select and use them in various situations.

The document was the product of successive revisions of an original draft prepared by the Committee on Test Standards of the American Psychological Association. It represents the coöperative efforts of this original group, a committee of the American Educational Research Association, and a committee of the National Council on Measurements Used in Education. These professional bodies constitute the major measurement groups in the United States and no document in any field could have a more authoritative source or sponsorship.

DEVELOPMENT AND SCOPE

The history of psychological testing covers a span of some fifty years or less. In that time, however, thousands of tests have been produced. *Some of the variability in quality of tests may exist because the field of testing is still in relatively early stages of exploration and experimentation.* Part might be accounted for by hasty attempts to meet competition in a highly competitive sales market, and some of it can certainly be attributed to the lack of specific standards backed by strong professional support.

While there has long been a general concern about the improve-

ment of testing among professional persons in the field, the idea of "standards" has met with some resistance. This resistance is based, at least in part, on the grounds that "standards" might inhibit innovation and experimentation in test construction. Some of this early philosophy of test development is reflected in the following statement made by a committee of the American Psychological Association in 1906: "Let many tests be tried, each new investigator introducing his own modification, and then, the worthless will gradually be eliminated and the fittest will survive."

Unfortunately, the worthless have not been eliminated. The staying power of some demonstrably inferior tests has been phenomenal. The members of the committee that prepared the recommendations in 1954 were not unaware of the fact that specifications for tests might discourage the development of new tests.⁴ They believed, however, that "appropriate standardization of tests and manuals need not interfere with innovation." The recommendations were intended by the committee to provide assistance to the producers of tests "to bring out a wide variety of tests . . . and to make those tests as valuable as possible."

The general principle or concept underlying the Technical Recommendations is that "a test manual should carry information sufficient to enable any qualified user to make sound judgment regarding the usefulness and interpretation of the test."

The recommendations are significant in that they suggest, directly or by implication, the kinds and quality of data that must be gathered before a test is released for use. They emphasize the fact that a test manual should leave the test user with an accurate impression of the test that goes beyond literal truthfulness. Test manuals must be written in such a way that those with limited training will not get a distorted idea of what the test will do.⁵ At

⁴ *Technical Recommendations, op. cit.*, p. 1.

⁵ The reader might turn back to Chapter III to reexamine the quotations from the stated purposes of a variety of tests and consider the impression such statements might leave on those who have not reached a high degree of sophistication about tests.

the same time, they "should be sufficiently complete for specialists in the area to judge the technical adequacy of the test."

In preparing the standards, no attempt was made to set statistical specifications as they relate to validity and reliability coefficients, standardization, or other quantitative aspects of tests but the need for enough data on such factors to permit judgment of their adequacy is stressed.⁶ The user must assume the responsibility for estimating the adequacy of the data presented before he employs it. He must decide whether they are sufficient in quality and quantity.

The standards presented in the document are intended to apply to virtually the whole range of measurement instruments, from achievement, ability, and aptitude tests through interest and personality inventories, projective instruments, and related clinical techniques. While some standards apply to all such devices, others are rather specific to particular types. In this respect the document recognizes several levels of test development. The highest degree of development, it is pointed out, is needed when tests are used in "practical" situations, where the user cannot obviously validate the test for his own purposes and must rely on the manual for data supporting the stated purpose and uses of the test. The recommendations are directed primarily to tests falling in this category. These are the type most frequently employed by the counselor. Not all tests, of course, are of this type. The effective use of some types, such as the projective tests, are dependent upon the clinical interpretation of qualitative responses. Arguments to the contrary notwithstanding, the document holds that these devices, too, should be accompanied by appropriate evidence about validity, reliability, and other factors related to test interpretation.

LEVELS OF RECOMMENDATIONS

The Technical Recommendations is essentially a listing of 165

⁶ *Technical Recommendations*, op. cit., p. 2.

statements, each representing a standard related to some particular aspect of test presentation. The standards are subsumed under the general topics of Dissemination of Information, Interpretation, Validity, Reliability, Administration, and Scales and Norms. Each standard is further classified in terms of its relative implications for the operational use of a test. Thus, the individual standards are designated as *essential*, *very desirable*, or *desirable*, in importance. The categories are defined in the document as follows:

The ESSENTIAL standards indicate what information will be genuinely needed for most tests in their usual applications. When a test producer fails to satisfy this need, he should do so only as a considered judgment. In any single test, there will be very few ESSENTIAL standards which do not apply. . . . A test manual can satisfy all the ESSENTIAL standards by clear statements of what research has and has not been done and by avoidance of misleading statements. The category VERY DESIRABLE is used to draw attention to types of information which contribute greatly to the user's understanding of the test. They have not been listed as ESSENTIAL for a variety of reasons. For example, if it is very difficult to acquire information (e.g., long-term follow-up), it can not always be expected to accompany the test. At times a closely reasoned minority opinion regards a type of information as unimportant. Such information is still very desirable, since many users wish it, but it is not classed as ESSENTIAL so long as its usefulness is debated.

The category DESIRABLE includes information which would be helpful, but less so than the ESSENTIAL and VERY DESIRABLE information.²

The application of the standards, reflecting each of these categories as applied to one of the topics, Dissemination of Information, is presented in the following examples. It is regarded as *essential*, for instance, that:

A2.2 When a test is revised or a new form is prepared, the manual

² *Ibid.*, pp. 5-6.

should be thoroughly revised to take the changes in the test into account.⁸

Furthermore, it is regarded as *very desirable* that:

A2.21 When a short form of a test is prepared by reducing the number of items or organizing a portion of the test into a separate form, new evidence [should] be obtained and reported for that new form of the test.⁹

It is *desirable* that:

A2.22 When a short form is prepared from a test, the manual [should] present the correlation between the long and short forms, separately administered.¹⁰

There would be little debate about the fact that a revision of a test, implying new items, would necessitate treatment of standardization, validity, reliability, and norms as thoroughly as that required for a new test. This being the case, many of the data in an existing manual would not apply to the revision and thorough reworking of the manual would be *essential*. In general, it could be reasoned that a short form of a test is a new test and hence it would be equally *essential*, rather than just *very desirable*, that the manual be revised. Within the framework of the document's definition of levels, however, the placing of the second standard in the category of *very desirable* may be appropriate. It is reasonable, too, that if sufficient data are presented for a short form of a test so it can stand alone, it would not be essential that the correlation between the long and short form be presented. Knowledge of the degree to which the two correlate, however, might be helpful in making a decision as to which form to use in specific instances.

Should the reader feel that the definitions of *very desirable* and *desirable* represent too much of a compromise, he may allay his fear in part by noting Table 19 below. Nearly three fourths of the

⁸ *Ibid.*, p. 9.

⁹ *Ibid.*, p. 9.

¹⁰ *Ibid.*, p. 9.

Measurement for Guidance

TABLE 16. Distribution of Standards by Topic and Category Level *

Standards Topic	Category			Not Categorized	Total
	Essential	Very desirable	Desirable		
A. Dissemination of Information	8	1	1		10
B. Interpretation	14	4			18
C. Validity (general)	5				5
Content	7				7
Predictive	31	13	2		46
Concurrent	2	1			3
Construct	1	4	1		6
D. Reliability (general)	11	3	1		15
Equivalence of forms	2	1			3
Internal consistency	6	1			7
Stability	4	2			6
E. Administration and Scoring	6	2	1	1	10
F. Scales and Norms	17	8	4		29
Total	114	39	11	1	165

* Source: *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Supplement to the *Psychological Bulletin*, March, 1954.

total of 165 standards are categorized as *essential*. Those categorized as *very desirable* (which cannot be too easily ignored by test authors and publishers in the future), in combination with *essential*, account for over 90 per cent of the total.

It will be noted further that more than one third of the standards relate to validity, the aspect of tests of most concern to the counselor. By far the greatest demands are placed on standards surrounding predictive validity, required most often in the counseling situation.

LIMITATIONS

While the Technical Recommendations represent an important milestone in the history of measurement, it must not be thought that all that remains is to have authors and publishers conform. The present document does not represent the final word, nor does it represent a present ideal. It has limitations.

Some of the limitations are those resulting from the very nature of psychological testing as a field. It is unlikely that the last word can ever be written in a field based on continuous research and experimentation and which produces new concepts, products, and applications.

Other limitations are due, in a sense, to human factors. In spite of the fact that the committees authoring the document represent high authority in the field of testing, it is unlikely that this group or any group could anticipate all aspects of testing or the questions and exigencies arising from the use of tests.

Still other limitations relate to the element of compromise, both of professional opinion and between the ideal and the practical.

These limitations are acknowledged in the document itself: "Despite the care with which the standards have been developed, experience will no doubt reveal that some of our judgments would benefit from further examination. New tests will present problems not considered in the present work. The improvement of statistical techniques and psychometric theory will yield better bases for test analysis. The efforts of test producers will lead to continued improvement in tests, and as this continues it will be possible to raise the standards."¹¹

Compromise is almost always inevitable when several groups attempt to evolve standards on which all can agree. That compromise played a part in the preparation of the Technical Recommendations is evident in the following: "In arriving at those requirements (as to information accompanying published tests), it has been necessary to judge what is presently the reasonable degree of compromise between pressures of cost and time, on the one hand, and the ideal on the other."¹² This compromise is reflected in part by the levels that were evolved for the various standards, and is noted particularly in the explanation of the level termed very desirable: "At times, a closely reasoned minority opin-

¹¹ *Ibid.*, p. 7.

¹² *Ibid.*, pp. 2-3.

ion regards a type of information as unimportant. Such information is still very desirable, since many users wish it, but it is not classed as ESSENTIAL so long as its usefulness is debated."¹³

Not all that is known at this time about good test practices is included in the present standards. The standards related to the administration of tests, for instance, are minimal. If followed literally and exclusively in a test manual, the test user would find that many questions related to administrative practices remained unanswered.¹⁴

That the standards are not as stringent as present knowledge could have them is suggested in another statement in the document: "Ideally manuals should be tested in the field by comparing typical readers' conclusions with the judgment of experts regarding the test. In the absence of such trials, our recommendations are intended to apply to the spirit and tone of the manual as well as its literal statements."¹⁵

One further limitation relates to the problem of enforcement. This is not a limitation of the document itself, for there is obviously no way to build enforcement into it, but is a problem of considerable importance. A proposal that a "Bureau of Test Standards" be planned in connection with Technical Recommendations was not favorably received at the time the American Psychological Association Committee on Test Standards was set up. The standards as presented are intended to be used without reference to enforcement machinery. It is unfortunate, in one sense, that the consumer is not protected in the area of psychological testing as well as he is in the area of patent medicines and drugs. The test user, then, must still screen the tests he uses. The Technical Recommendations will help him do the job.

The Coöperative School and College Ability Tests were among

¹³ *Ibid.*, p. 6.

¹⁴ Some test manuals, such as that of the *School and College Ability Tests* reviewed in this chapter, far exceed the demands regarding test administration as presented in the *Technical Recommendations*.

¹⁵ *Ibid.*, p. 2.

the first tests to be published after the recommendations were published. The remainder of this chapter is devoted to an evaluation of the Manual prepared to accompany this series. The standards presented in the Technical Recommendations appropriate to this type of test will be applied to determine the adequacy of the test. The procedure used in the following pages is one that counselors might well employ before they purchase a test.

APPLICATION OF RECOMMENDATIONS TO THE COOPERATIVE SCHOOL AND COLLEGE ABILITY TESTS¹⁶

The tests in the SCAT series are designed to help "teachers and counselors—and students themselves—to estimate the capacity of each *individual student* to undertake the academic work of the next highest level of schooling." The series consists presently of tests suitable for use at five levels: college freshmen (Level I), senior high school (Level II), Grades 8, 9, and 10 (Level III), upper elementary (Level IV), and intermediate elementary (Level V). A sixth level, suitable for superior college sophomores, is tentatively planned.

The principal objective of the series, as stated in the Manual, is to provide continuity of measurement over a long range of years, extending from about the fourth grade through the sophomore year in college. The series, according to the Manual, should "make it possible to chart and study the growth of individual students over a range of years not now possible."

The SCAT tests were developed as an alternative to revision of the American Council on Education Psychological Examinations. Like the older ACE, the SCAT yields scores in verbal and quantitative areas and a total representing the combined scores.¹⁷ The content and approach of the new test differ from that of the ACE,

¹⁶ *The Cooperative School and College Ability Tests*, Princeton, N.J., Cooperative Test Division, Educational Testing Service, 1955.

¹⁷ *Ibid.*, p. 5.

however, since the SCAT attempts to get at school-learned abilities that are "critical prerequisites to next steps throughout the range of general education." These abilities are described as "comprehending the 'sense' of a sentence read, attaching meaning to isolated words, manipulating numbers and applying number concepts accurately in a computation situation, and solving quantitative problems."

The publishers have announced that they intend to "encourage the use of SCAT as the best we have to offer for measurement of academic ability." They recognize, however, that immediate change-over from the ACE will be neither practical nor desirable in some cases and they plan to make the ACE available at least until 1959. To aid in the transition, the publishers have equated the most recent college and high school editions to the SCAT score scale. This will make it possible for schools to change over, if they desire to do so, without loss of valuable local norm data on the ACE that may have been collected over a long period.

Some fifteen uses for the tests as they relate to teaching, counseling, and administration are listed on page 2 of the Manual. Those related specifically to counseling follow:

- . . . when the tests are used for their principal purpose, the counselor can apply the results in his work with students to:
- a. help the student to understand his own strengths and weaknesses in comparison with students in certain norming groups;
 - b. guide the student toward choices of educational goals and courses most appropriate for him;
 - c. estimate the levels of achievement to be expected of the student;
 - d. compare the measured academic abilities of students in different class, grade, and school groups.

The extent to which these suggested uses are supported by appropriate validity data as well as the extent to which the Manual meets other standards of the Technical Recommendations will be noted in the analysis below. It is hoped that the method of analysis

applied here may be useful to those counselors who are required to select tests.¹⁸

A. DISSEMINATION OF INFORMATION

TECHNICAL RECOMMENDATION

A1. When a test is published for operational use, it should be accompanied by a manual which takes cognizance of the detailed recommendations in this report. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This initial recommendation appears to be met. Direct reference to the Technical Recommendations is made in the discussions of validity and reliability. The Manual points out clearly the limitations of the instrument. It contains discussions (page 9 of Manual) of "Inferences that validity data and arguments are intended to support," "Inferences of the kind which are at present supported only by *assumption* of validity," and "Inferences of the kind that are *not* supported by any evidence of validity, and which the user of the tests should avoid in every case." Such statements as "The usefulness of tests in the SCAT series as predictors of future school or college success is as yet only assumed," further attest to the efforts of the publishers to meet the technical recommendations through acknowledgment of areas of weakness and limitations.

TECHNICAL RECOMMENDATION

A1.1 Some form of manual, presenting at least minimum informa-

¹⁸ The letter-number code that appears in the paragraphs of the analysis are the ones used in the *Technical Recommendations* to designate the specific standards quoted from the document. There are gaps in the letter-number designations because all the standards do not apply to tests of this type.

tion, should be given or sold to all purchasers of the test. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

A Manual is provided in which data pertinent to the test are presented in great detail. As with some other tests marketed in recent years, the Manual is not furnished with the test but must be bought separately. This seems to be justified since the 57-page Manual represents more expense to the publisher than the very brief and inadequate manual commonly offered. The tests themselves are accompanied by 14-page pamphlets containing directions for administration, scoring, and discussion of the interpretation of individual scores and group interpretations.

TECHNICAL RECOMMENDATION

A1.2 Where the information is too extensive to be fully reported in such a manual, the manual should summarize the ESSENTIAL information and indicate where further details may be found. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

As indicated in A1.1 above, the Manual itself presents pertinent information in considerable detail. Since the test is new, no body of information regarding the tests, other than that reported in the Manual, has developed up to the time of this writing.

TECHNICAL RECOMMENDATION

A2. The manual should be up-to-date. It should be revised at appropriate intervals. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT series, at this writing, is new and is accompanied by

the initial Manual. The user of the test is promised, however, that there will be at least one revision of the Manual during the first two years of its use. Several supplements containing additional norm data and research findings are also planned. One supplement appeared in 1958. The purchaser is advised that these revisions and supplements will be sent to him automatically and *without charge*. A tearout postal card is provided containing this statement: "Return of this service card, properly filled in, entitles you to continuous and automatic supplement service through 1960 without additional cost. Please send it now." This approach, used also in a few earlier tests, is commendable—a practice that all test publishers would do well to follow.

TECHNICAL RECOMMENDATION

A2.1 When new information emerges from investigations by the test authors or others, which indicates that some facts and recommendations presented in the manual are substantially incorrect, a revised manual should be issued at the earliest feasible date. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This recommendation, at present, is not relevant. The promise of a new, revised manual within two years and periodic supplements containing additional research data, however, suggests that it is the intent of the publisher to keep the user abreast of the latest data, and to satisfy this recommendation.

TECHNICAL RECOMMENDATION

A2.3 The copyright date of the manual or the date of the latest revision should be clearly indicated. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This standard is met. Two supplements containing additional norm data are similarly dated.

tion, should be given or sold to all purchasers of the test. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

A Manual is provided in which data pertinent to the test are presented in great detail. As with some other tests marketed in recent years, the Manual is not furnished with the test but must be bought separately. This seems to be justified since the 57-page Manual represents more expense to the publisher than the very brief and inadequate manual commonly offered. The tests themselves are accompanied by 14-page pamphlets containing directions for administration, scoring, and discussion of the interpretation of individual scores and group interpretations.

TECHNICAL RECOMMENDATION

A1.2 Where the information is too extensive to be fully reported in such a manual, the manual should summarize the **ESSENTIAL** information and indicate where further details may be found. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

As indicated in A1.1 above, the Manual itself presents pertinent information in considerable detail. Since the test is new, no body of information regarding the tests, other than that reported in the Manual, has developed up to the time of this writing.

TECHNICAL RECOMMENDATION

A2. The manual should be up-to-date. It should be revised at appropriate intervals. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT series, at this writing, is new and is accompanied by

the initial Manual. The user of the test is promised, however, that there will be at least one revision of the Manual during the first two years of its use. Several supplements containing additional norm data and research findings are also planned. One supplement appeared in 1958. The purchaser is advised that these revisions and supplements will be sent to him automatically and *without charge*. A tearout postal card is provided containing this statement: "Return of this service card, properly filled in, entitles you to continuous and automatic supplement service through 1960 without additional cost. Please send it now." This approach, used also in a few earlier tests, is commendable—a practice that all test publishers would do well to follow.

TECHNICAL RECOMMENDATION

A2.1 When new information emerges from investigations by the test authors or others, which indicates that some facts and recommendations presented in the manual are substantially incorrect, a revised manual should be issued at the earliest feasible date. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This recommendation, at present, is not relevant. The promise of a new, revised manual within two years and periodic supplements containing additional research data, however, suggests that it is the intent of the publisher to keep the user abreast of the latest data, and to satisfy this recommendation.

TECHNICAL RECOMMENDATION

A2.3 The copyright date of the manual or the date of the latest revision should be clearly indicated. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This standard is met. Two supplements containing additional norm data are similarly dated.

B. INTERPRETATION

TECHNICAL RECOMMENDATION

B1.1 Names given to tests, and to scores within tests, should be chosen to minimize the risk of misinterpretation by test purchasers and subjects. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The test title, *School and College Ability Tests*, and explanations of intent are evidence that this recommendation has been met. Frequent mention is made of the basic purpose of the instrument—that of “helping teachers and counselors—and students themselves—to estimate the capacity of each individual student to undertake the academic work at the next higher level of schooling. The tests are measures of developed ability, indicative of the relative academic success the student is likely to achieve in his next steps up the educational ladder.”¹⁹ Considerable discussion is presented regarding “school-learned abilities.” Those regarded as most critical for the purpose at hand are clearly named and were used in evolving the final items. Care too has been taken in pointing out that the test scores do not indicate the ‘intelligence’ or ‘native capacity of the student,’ that the abilities measured are not to be interpreted as fixed or permanent characteristics of the student, and that the scores on the tests do not, at this time, indicate an individual’s likelihood of success in vocational training or in certain occupations. Further compliance with this recommendation is found in the Manual’s discussion of “Interpretation of Individual Scores and Group Distributions” (p. 32 App. A). “The Coöperative School and College Ability Tests are intended to measure four of the school-learned skills which research has shown to be closely related to academic success in school and college. These tests are NOT measures of ‘intelligence’ in the sense that they

¹⁹ *Ibid.*, p. 3.

tap innate psychological characteristics, nor are they really measures of 'aptitude' because aptitudes usually are regarded as fairly stable characteristics not greatly affected by instruction."

This statement, unfortunately, is somewhat buried in the Manual. It might well have been given a more prominent place, most desirably in the opening discussion of purposes and uses. In another section related to interpretation, the user is cautioned against overinterpretation (p. 20).

Even though the tests in the SCAT series have been constructed to yield scores of high reliability and the directions are quite specific as to what is measured and what is not measured, users of the tests are cautioned *not to over-interpret the test scores*. *The tests measure well the things they are intended to measure*, but educators who are not extensively trained in educational and psychological measurement often are tempted to "read into" their interpretation of any test scores some conclusions which the scores will not support. For this reason, the section of the "Directions" which contains suggestions on interpretation specified some of the things the tests do *NOT* measure. . . .

Thus the user is cautioned frequently in the Manual regarding interpretation and limits thereof. Here again the publishers must be commended for the straightforward manner in which the purposes and scores of the tests are presented. There appears to be no attempt directly or by implication to represent the instrument as something it is not.

TECHNICAL RECOMMENDATION

B1.2 The manual or other accompanying material should describe *the process by which interpretations are to be derived from test scores*.
VERY DESIRABLE.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

If one could assume that tests would be used only by those who

automatically consider other data in addition to test scores there would be no argument about the present practice of suggesting to the user that the instrument at hand is something of a panacea for all the frustrations, questions, and problems arising from working with other people in an educational or counseling situation. But this is a gross assumption. Considering the wide use and misuse of psychological tests, and the fact that they may be obtained and used by many who have no particular training or background in testing, it would appear *essential* that all test manuals include prominently a statement reminding the test user of the dangers of interpretation in isolation. They must encourage consideration of the factors that may have influenced the test performance and the types of data one might seek for evidences of contradiction. The SCAT, in terms of this standard, is only a slight improvement over those tests marketed prior to the appearance of the Technical Recommendations. There are a few token statements, however, that may suggest to the user that other factors must be taken into account. The following statement from page 8 of the Manual is made in the discussion of validity: "The test appears to be a 'work sample' measuring the ability of the students in several of the skills that are important to further academic learning. If these were the *only* abilities needed for success in school, and if the criterion of teacher-assigned grades were a reliable one, the validity coefficients of a highly reliable test would be very close to 1.00. There are many other factors also important to school success, however—other academic abilities as well as habits and attitudes. . . ."

That other factors need to be taken into account is also implied indirectly in the following discussion of norms, which appears on page 33 of Appendix A of the Manual.

A more useful set of norms can be built locally . . . and are more useful than "national norms" for most interpretations because they reflect the scores of students who are *much like the student now being tested* in a number of ways: (a) they are at about the same age in

each grade because they enter school and are promoted from grade to grade under a common system; (b) they are studying in the same curriculum; (c) they come from generally the same kinds of homes; (d) the over-all quality of their instruction is about the same; (e) the cultural and academic advantages offered by the community are similar; and (f) the school services available to them are the same. All of these things affect the test performance of students in some way. . . .

The SCAT user, then, is reminded, at least indirectly, of some of the factors that need to be taken into account in interpreting the test results. It would be desirable to have these points, and others, stated more directly under a separate heading, but their very presence in any connection is encouraging.

TECHNICAL RECOMMENDATION

B1.22 When case studies are used as illustrations for the interpretations of test scores, the examples presented should include some relatively complicated cases whose interpretation is not clear-cut.
VERY DESIRABLE.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This recommendation, as with the one previous, is regarded as *very desirable* rather than *essential*. In a sense, then, the Manual should not be *required* to meet this particular standard. Since the Manual does present an illustrative case, however, some comment seems appropriate. The case does not appear to be particularly complicated, and the interpretation is somewhat perfunctory, though it may aid the test user in some aspects of interpretation. It does tend to emphasize one of the strongest features of the Manual, the need to interpret test scores in terms of ranges and confidence intervals. Again, however, the illustrative case is confined almost exclusively to test results and fails to introduce non-test data that may be of consequence in counseling. The value of

the case would be greatly enhanced by having results integrated with other counseling data, since that is the way the results presumably would be used. The ultimate implications of test results are not discussed in any detail.

TECHNICAL RECOMMENDATION

B2. The test manual should state explicitly the purposes and applications for which the test is recommended. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The statement of purposes of the test presented in the Manual is typical of most such statements. It is interesting to note, however, that the Manual, in presenting the possible applications of the tests, suggests that they can "aid" the teacher, that the counselor can "apply" the results, and that the data from the test results can "help" the administrator. This is a refreshing contrast to many test manuals which state rather unequivocally that the tests *will do* certain things rather than indicate that they *may help* in doing them. The limitations implied by the verbs employed, however, are likely to be overlooked by many users. Greater emphasis could be given to this point. Though perhaps not wholly relevant here, it appears to be assumed that the test user will have the know-how required to use the results in the various ways suggested. It would be helpful to have the stated uses supplemented with examples.

TECHNICAL RECOMMENDATION

B3. The test manual should indicate the professional qualifications required to administer and interpret the test properly. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT tests, it is pointed out in the Manual, were "developed and arranged in the expectation that they will be adminis-

tered, scored, and interpreted mostly by teachers who have had little or no formal training in testing." It is believed by the authors that any teacher or counselor who takes time to study the directions and follow them will do as well as a test expert. This seems reasonable, of course, since a test expert could do no more. As indicated in E1 below, the treatment of directions for administering the test is very thorough. It is believed by the publishers that counselors and others with training in measurement should have little difficulty in using and interpreting the test.

Some control over use of the SCAT is attempted through limitations of purchase established by the publisher. "Administrators, college teachers, and professionally qualified advisors in recognized schools . . . through order on official letterhead or purchase order forms. High school teachers can similarly purchase the tests with written approval of their administrators, graduate students with approval of their instructors. Staff members of other organizations and individuals in private practice having a master's degree in psychology or education, or equivalent in training and experience can purchase upon presenting their qualifications. The company reserves the right to accept or reject orders . . . in conformity with professional standards." (Manual, page 57.)

TECHNICAL RECOMMENDATION

B3.11 The manual should not imply that the test is "self-interpreting," or that it may be interpreted by a person lacking proper training. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual does not imply in any sense that the test is self-interpreting. It does imply that teachers in general should be able to interpret the tests by reading the Manual. It is somewhat inconsistent, however, in suggesting, as above, that those with training

in measurement should have no difficulty in interpreting the test, but those "who have not acquired extensive training in measurement" should follow the Manual in "cookbook" fashion. While the manuals should be as complete as possible, it seems unlikely that they will ever be detailed enough so that a straight "cookbook" approach will be sufficient. Perhaps much of the misuse of tests can be attributed to the attempts on the part of test publishers to make tests simple and to imply that anyone who can read can also give tests. While more cautious about this point than the manuals of many other tests, the cause of testing might better have been served by urging that *all* users be trained in measurement.

TECHNICAL RECOMMENDATION

B3.12 The manual should point out the counseling responsibilities assumed when a tester communicates interpretations about ability or personality traits to the person tested. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The value of any test would be greatly enhanced if the manual would remind the user of the implications of the results, and that other data must be considered when interpreting results to an individual. The user must be reminded that test results should not be interpreted in isolation and that the student should not be left in a state of suspension. The counselor must be sure that the counselee is following his interpretation and that he has a grasp of percentile or other statistical concepts that may be used. In this regard, the SCAT Manual leaves room for improvement. If the authors assumed some of these points, this is not made clear in the Manual. It does not appear to contain any direct acknowledgment in the form of discussion of this standard. One of the stated counseling uses of the test, that of helping the students "to understand his own strengths and weaknesses . . .," implies interpretation

of results to the student. While no direct comment is made as to how this is to be done, there is nothing in the Manual that should lead the test user to believe students can or should make their own interpretations.

TECHNICAL RECOMMENDATION

B5. Statements in the manual reporting relationships are made by implication quantitative, and should be stated as precisely as the data permit. If data to support such a statement have not been collected, that fact should be made clear. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

What this standard appears to demand is some expression of the *degree* to which (in this case) further academic learning at a given level depends upon the abilities measured by the test. If one expected to find a statement in the Manual to the effect that "90 percent of academic success is due directly to the factors measured by this test," he will not find it. The problem of relationship of test scores to academic performances has long confronted test authors, counselors, teachers, registrars, and others. The SCAT does not appear to provide, at this time, any better solution to the problem than other tests. The authors do not, however, claim perfect correlation between SCAT scores and further academic learning. They state only that it measures "several skills that are important." The validity coefficients presented are not significantly higher than those obtained previously by use of other tests. Results of further studies in progress are promised to those who use the test as soon as they are known.

TECHNICAL RECOMMENDATION

B5.2 The manual should clearly differentiate between an interpretation justified regarding a group taken as a whole, and the application

of such an interpretation to each individual within the group. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

One of the strong features of this test is the emphasis placed on careful interpretation and the caution against overinterpretation. This is strengthened by the introduction of the "confidence interval" concept. The reader is urged to interpret scores as a "band of possible test scores" rather than as an exact point. The standard error of measurement is built into the scoring and recording procedure. It places more usable limits on individual score interpretation than those afforded by the coefficients of reliability that are usually presented in test manuals. The careful reader will note the explanation that the interpretation of this approach is that "the chances are two-to-one that the student's 'true score' lies within this interval." This point could be strengthened, of course, by a statement that there is still one chance in three that the student's true score would fall beyond the interval. This standard is acknowledged in another statement on page 11 of the Manual. "The user of the *School and College Ability Tests* cannot be reminded too emphatically that the scores of individuals on the four subtests should NOT be interpreted separately. The part scores and total scores . . . are reliable enough for individual use, but separate subtest scores should *never* be recorded for individuals." (*Italics theirs.*)

C. VALIDITY

TECHNICAL RECOMMENDATION

C1. When validity is reported, the manual should indicate clearly what type of validity is referred to. The unqualified term "validity" should be avoided unless its meaning is clear from the context. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT Manual, in presenting validity data, quotes directly from the Technical Recommendations and organizes its material around the four types of validity defined in the document as *content*, *predictive*, *concurrent*, and *construct*.

TECHNICAL RECOMMENDATION

C2. The manual should report the validity of each type of inference for which a test is recommended. If validity of some recommended interpretation has not been tested, that fact should be made clear.
ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The manner in which this standard is met by the SCAT Manual is, in general, good. The Manual presents, in a specifically titled section, the inferences that presumably are supported by data and expert opinion, those that are supported by *assumption* of validity, and inferences "of the kind which are NOT supported by any evidences of validity." The Manual cautions the user to avoid these latter in every case. The Manual offers three inferences that the arguments and data are intended to support. One is based on "expert opinion," one on estimated concurrent validity coefficients, and one on a combination of the two. It is commendable that the Manual makes a clear distinction between the inferences. It is well, too, that the test was conceived with the help of much expert opinion. This makes a good starting point. It is unfortunate that the authors of the test, at the time of its introduction, did not present more adequate statistical evidence in support of the inferences presumably based on the expert opinion and on concurrent validity coefficients.

TECHNICAL RECOMMENDATION

C2.1 The manual should indicate which, if any, of the interpretations usually attempted for tests such as the one under discussion have not been substantiated or are based merely on clinical impressions. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

In addition to advising the user regarding inferences that the test is and is not intended to support, the Manual clearly states that predictive validity "is as yet only *assumed*" and that concurrent validity, as reported, is estimated (on the basis of experimental tests) and that further studies are under way. In the light of these statements, this standard is adequately met.

TECHNICAL RECOMMENDATION

C3. Findings based on logical analysis should be carefully distinguished from conclusions established by correlation of test behavior with criterion behavior. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

There appears to be no attempt to present findings based on logical analysis alone, that is, where items representing specific areas were assembled and validity assumed on the efficacy of the items themselves. The evidence of content validity is rather based on a combination of factors (including concurrent validity) as noted in the following statement:

Thus, by logical and empirical means, there were developed and tried out a series of practical measures of school-learning ability having the following characteristics in content validity:

- a. Measurement of developed abilities rather than innate psychological traits;
- b. Measurement of abilities which a committee of noted educational

researchers recommended as most closely related to success in school learning;

- c. "Face validity" in the sense that students and parents can see in the test content a measurement of abilities related to school learning;
- d. Relatively high correlations with school marks assigned by teachers;
- e. Sufficiently low intercorrelations between the scores on the verbal and quantitative parts of the test to indicate measurement of somewhat different abilities.²⁰

Logical analysis was employed in the initial selection of the experimental test content, but conclusions regarding content validity are based heavily on the degree to which test types correlated with class grades in the areas they were supposed to measure. None of the content validity data is based on the final form of the test. This point is made on page 6 of the Manual as follows: "Although these data, obtained with experimental forms of the test, are the only concurrent validity coefficients offered at the time of publication, it is likely that they are *minimum* estimates." (*Italics theirs.*)

It is unfortunate that the data are not based on the final form of the test, that the user does not have *actual* rather than *estimated* coefficients, and that he does not have some evidence that they are minimum. At the same time, it is commendable that much of the discussion of content validity data is couched in tentative rather than positive terms where data are lacking. If the reader will heed such statements as "these characteristics *suggest* that the test *probably* will have sufficiently high content validity to make it *useful* . . ." (*italics added*) he will perhaps exercise appropriate caution when using the test and save his enthusiasm for the time when more positive evidence becomes available.

TECHNICAL RECOMMENDATION

C4. If a test performance is to be interpreted as a sample of perform-

²⁰ *Ibid.*, p. 8.

ance in some universe of situations, the manual should indicate clearly what universe is represented and how adequate the sampling is. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This standard is most applicable to tests of achievement or proficiency in some subject area, e.g., mathematics, reading. The SCAT is not intended to be a test of this kind. At the same time, almost all tests are samples of a performance in some universe of situations and in this case the universe appears to be "school-learned abilities" or skills. The four subtests comprising the final form of the test were screened by statistical and logical analysis from nine "abilities" recommended for such screening by the advisory committee. These nine abilities constituted in one sense the "universe" sampled by the final test. They included Resourceful Computation, Reading Comprehension, Sentence Completion, Analogies, Routine Computation, Data Sufficiency, Vocabulary, Mixed Computation, and Arithmetic Reasoning. While certain performances may be inferred by these test types or abilities, no specific description of their content is offered. Since the final form of the test consists of only four of the nine abilities (Routine Computation, Arithmetic Computation, Sentence Completion, and Vocabulary), only a partial sample appears to have been taken. This limitation of sampling is recognized in the Manual, however, and the test is not represented as measuring *all* the school-learned abilities needed in further academic learning.

TECHNICAL RECOMMENDATION

C4.1 The universe of content should be defined in terms of the sources from which items were drawn, or the content criteria used to include and exclude items. **ESSENTIAL.**

C4.2 The method of sampling items within the universe should be described. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Here again the standard applies most obviously to tests of subject matter achievement or proficiency, but has some bearing on the SCAT. The Manual indicates only that "test items were designed to measure these nine abilities and a test of each type was built for experimental use" (Manual, p. 5).

A 30-item vocabulary subtest is included in the final form, but no mention is made in the Manual as to the source of vocabulary items. Arithmetic reasoning and routine computation items are used in two of the subtests, but again, neither the source nor the criteria for selecting the items are presented. The reader must *assume* that these items are the most suitable for measuring the "abilities" that the test is designed to sample. The criteria used to select the best combination of test *types* for inclusion in the final form of the test are presented in the Manual.

TECHNICAL RECOMMENDATION

PREDICTIVE VALIDITY. Of 67 recommendations relating to all types of validity, 47 related specifically to the predictive type. They are designed also to cover concurrent validity. Because the SCAT manual does not present evidence of predictive validity (see accompanying comment), no direct analysis can be made. Some of the recommendations relevant to predictive validity are presented in the analysis which follows.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

No direct evidence of predictive validity is offered the SCAT user in the first Manual. This is the case in spite of the fact that one of the major purposes of the test (Manual, p. 3) is predictive in nature—"to estimate the capacity of each *individual student* to undertake the academic work of the next higher level of school-

ing." The absence of predictive validity data precludes at least one of the suggested counselor uses, to "guide the student toward choices of educational goals and courses most appropriate for him" (Manual, p. 2). This is not to say, however, that the Manual ignores the need for such data. It acknowledges the problem of securing data on predictive validity in the following paragraph on page 8.

The usefulness of tests in the SCAT series as predictors of future school or college success is as yet only *assumed*. It is reasonable to expect that measures of this type having respectable concurrent validity will also predict individual success well enough to be useful for prognostic purposes, for tests of these kinds have proved to be predictive in other forms and uses, but data on predictive power must be collected over a period of time and the prediction studies of these tests have not been completed at the time the series is first published. Prediction studies were initiated at each of several grade levels as soon as the final content of the tests had been determined; the results of these studies will be added to the manual in supplements as soon as they are known. Predictive validities for varying periods of time—from eight months to four or five years—and for different educational criteria will be reported.

Promises of things to come are of little comfort to the user who needs the evidence *now* if he is to use the test for predictive purposes. While one can deplore the absence of the data, the authors of the Manual must be commended for their treatment of the point. The need and absence are at least pointed out to the user. The authors could have as easily ignored the point, hoping that the users would, too.

CONCURRENT VALIDITY. On page 19 of the Manual one finds the following statement regarding concurrent validity:

The validity coefficients presented in the section on "content validity" are *estimated* coefficients of concurrent validity. That is, they are the statistically combined concurrent validities of the experimental test types that were used to determine the final forms in the series.

Although these estimates can be regarded as accurate and useful for most practical purposes, further studies of the concurrent validity of the final forms are under way. The results of these studies, confirming and extending the original data, will be reported in an early supplement to the manual.

Thus the user is asked to accept the evidence of concurrent validity on the experimental tests. While it is probable that further studies on the final form of the test will confirm the findings of the experimental tests, it is unfortunate that the user does not have *actual* evidence to help him estimate the usefulness of the tests. In reading the following analysis of concurrent validity, the reader will keep in mind that the data are based on *estimated* coefficients.

TECHNICAL RECOMMENDATION

C5.1 Statistical procedures which are well known and readily interpreted should be used in reporting validity whenever they are appropriate to the data under examination. Any uncommon statistical techniques should be explained. ESSENTIAL.

C5.11 Reports of statistical validation studies should ordinarily be expressed by: (a) correlation coefficients of familiar types; (b) description of the efficiency with which the test separates groups, indicating amount of misclassification or overlapping; (c) expectancy tables. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Coefficients of correlation between test scores and grades are presented, but the method of computation is not described. It is probable that the usual Pearson correlation technique was employed, but this can only be assumed.

TECHNICAL RECOMMENDATION

C5.2 An over-all validity coefficient should be supplemented with

evidence as to the validity of the test at different points along the range, unless the author reports that the validity is essentially constant throughout. VERY DESIRABLE.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This standard is not met by the SCAT Manual.

TECHNICAL RECOMMENDATION

C5.3 Test manuals should not report coefficients corrected for unreliability of the test as estimates of predictive validity. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

There is no evidence that the reported coefficients have been corrected for attenuation.

TECHNICAL RECOMMENDATION

C6. All measures of criteria should be described accurately and in detail. The manual should evaluate the adequacy of the criterion. It should draw attention to significant aspects of performance which the criterion measure does not reflect and to the irrelevant factors which it may reflect. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The criteria used in determining validity coefficients were total grade averages, English grades and mathematics grades in Grade 9, total grades, and English grades in Grade 12. The coefficients were computed from scores obtained from students in 19 high schools located in eight different states. The schools were divided into "High," "Medium," and "Low" categories on the basis of per-pupil investment in education by the community. The actual

differences between these categories are not described. The Manual does not offer a description of the size of the schools, their general philosophy, or objectives, nor is the nature of the classes in English or mathematics presented. The test user will have few data upon which to base judgment as to the representativeness of the schools included.

TECHNICAL RECOMMENDATION

C6.5 The time elapsing between the test and determination of the criterion should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The exact time lapse is not given in the Manual, though it is stated that the validity coefficients are based on grades earned during the same semester in which the experimental tests were given. It seems unlikely that short lapses of time would affect the results greatly in a test of this kind, as might be the case in achievement tests where much learning may take place during the time lapse.

TECHNICAL RECOMMENDATION

C7. The reliability of the criterion should be reported if it can be determined. If such evidence is not available, the author should discuss the probable reliability as judged from indirect evidence. **VERY DESIRABLE.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The reliability of school marks used in the determination of the validity coefficients is not reported, but their unreliability is conceded. The determination of reliability in each of the schools would in itself be something of a major undertaking. It is unlikely that such data, if gathered, would alter the generally known fact

that grades are unreliable. The Advisory Committee's recommendation that the test should measure school-learned abilities was based in part on the observation "that the best single predictor of how well a student is likely to succeed in his school work next year is how well he is succeeding this year" (Manual, p. 5). Grades, however unreliable in general, appear to be our best evidence as to how well the student is doing in school.

TECHNICAL RECOMMENDATION

C8. The date when validation data were gathered should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This is given only as 1953. As above, validation date on a test of this type is not likely to change radically over a period of a year or so, but it might over a longer period as curricular changes are made.

TECHNICAL RECOMMENDATION

C9. The criterion score of a person should be determined independently of his test score. The manual should describe precautions taken to avoid contamination of the criterion or should warn the reader of any possible contamination. **ESSENTIAL.**

C9.1 When the criterion consists of a rating, grade, or classification assigned by an employer, teacher, psychiatrist, etc., the manual must state whether the test data were available to the rater or were capable of influencing his judgment in any way, e.g., indirectly through other reports of the psychologist. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual is careful to note that grades were assigned to the

students involved in the validation study *before* the teachers knew their test scores.

TECHNICAL RECOMMENDATION

C13. The validation sample should be described sufficiently for the user to know whether the persons he tests may properly be regarded as represented by the sample on which validation is based. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

See C6 above.

TECHNICAL RECOMMENDATION

C13.1 The user should be warned against assuming validity when the test is applied to persons unlike those in the validating sample. **ESSENTIAL.**

C13.3 The number of cases in the validation sample should be reported. The group should be described in terms of those variables known to be related to the quality tested: these will normally include age, sex, socioeconomic status, and level of education. Any selective factor which restricts or enlarges the variability of the sample should be indicated. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Such a warning is not given directly, but may be implied in a discussion of the desirability of local norms. After listing the ways in which students may differ in terms of educational opportunity the following statement (C13.3) is made on page 33 of Appendix A of the Manual.

"All these things affect the test performance of students in some way, so that to estimate how well a student is doing *with what he*

has to start with it is important to compare him with others who have approximately the same start." This is a rather hollow standard in the case of the SCAT, however, for the persons in the validating sample are not described by age, range, or sex, and the implications of levels of "investment in education by the community" are not indicated. The number of cases is reported.

TECHNICAL RECOMMENDATION

C17. Reports of concurrent validity should be so described that the reader will not regard them as establishing predictive validity. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The discussions of the various types of validity are handled separately in the organization of the Manual. Some confusion does result from the inclusion of some concurrent validity data with that of content validity. Concurrent and predictive validity data, however, are adequately separated.

CONSTRUCT VALIDITY. Construct validity as a type is acknowledged on page 9 of the SCAT Manual with the following statement: "Since the comparative usefulness and accuracy of tests like those in the SCAT series can be demonstrated in terms of concurrent and predictive validity data—using criteria, which are varied, objective, and specific—there is no real need to investigate their construct validity. Studies of construct validity by interested users of the tests will be welcomed by the publisher as sources of interesting information about the tests, but no such studies will be undertaken to prove the worth of the instruments."

Considering the nature of construct validity, the publisher's arguments appear reasonable. Since the test is designed to replace the American Council on Education Psychological Test, however, it would be of interest to know how the two correlate. Such com-

parisons would fall within the scope of construct validity. No such comparisons are presented.

D. RELIABILITY

TECHNICAL RECOMMENDATION

D1. The test manual should report such evidence of reliability as would permit the reader to judge whether scores are sufficiently dependable for the recommended uses of the test. If any of the necessary evidence has not been collected, the absence of such information should be noted. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual of the SCAT is unusually thorough in this respect, reporting reliability estimates coefficients of internal consistency and coefficients of equivalence, with a good deal of empirical evidence. A careful statement is made that no coefficients of stability are available at the time of printing, but that such studies are under way and will be reported at a later date in supplements to the Manual.

TECHNICAL RECOMMENDATION

D1.1 Recommendation D1 applies to every score, subscore, or combination of scores whose interpretation is suggested. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

A special note is given in boldface type on page 11 of the Manual of the SCAT. It warns the test user not to forget that the *scores of individuals on the four subtests should NOT be interpreted separately for individuals*. The profile sheet does not allow room for the recording of these part scores to emphasize further the need to avoid this common error.

TECHNICAL RECOMMENDATION

D1.2 If differences between scores are to be interpreted or if the plotting of a profile is suggested, the manual should report the reliability of differences between scores. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The special construction of the test score profile sheet for the SCAT features a "confidence interval," which is portrayed by shaded lines so as to make it distinctive to the viewer. The following statement is made both in the Manual (page 11) and on the test profile sheet: "Use of the confidence interval makes it possible to note educationally important differences between the two scores (verbal and quantitative) at a glance. If the confidence intervals of the verbal and quantitative scores of an individual overlap, there is probably no educationally significant difference between the two scores and for interpretative purposes you can regard them as being equal. If the confidence intervals around the verbal and quantitative scores of an individual do NOT overlap, the chances are about 5-to-1 that an educationally important difference exists between the two scores."

TECHNICAL RECOMMENDATION

D1.5 Reports of reliability studies should ordinarily be expressed in terms of: (a) the product-moment correlation coefficient; (b) another standard measure of relationship suitable to categorical judgments; or (c) the standard error of measurement. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Both coefficients and error estimates are reported in the SCAT Manual. The reliability estimates for all subtests, the two major parts (verbal and quantitative) and total scores are all reported.

The Kuder-Richardson reliability coefficients are in the range of .82 and .88 for the part scores, and the Manual warns accordingly that these are not sufficiently high for use in individual test score interpretations. In a table reporting this information the standard error of measurement of each subtest, part, and total scores is also reported. This is given in raw score points and ranges from 2.0 for some of the subtests to 4.3 raw score points for the total score. This information is provided for both the high school "School Ability Test. Form 2A" and the college level "College Ability Test. Form 10." A sample of 370 cases was used with each reliability study. The total score reliability estimate is .95, which is regarded, quite properly, as being high enough to use with an individual case. It is in individual cases that counselors must seek some degree of confidence that the present score represents a close approximation of his counselee's true score.

TECHNICAL RECOMMENDATION

D2. The manual should avoid any implication that reliability measures demonstrate the predictive or concurrent validity of the tests. ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The section on reliability in the Manual of this test is introduced almost word for word with the general presentation of the topic of reliability in the Technical Recommendations. After this general discussion, various types of reliability information are presented. Nothing is presented in this section that might lead the test user astray.

TECHNICAL RECOMMENDATION

D3. In reports of reliability, procedures and samples should be described sufficiently for the reader to judge whether the evidence

applies to the persons and problems with which he is concerned.
ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Coefficients of internal consistency are reported as the best type of evidence of reliability of this test. Information is reported for both the high school form, SAT, and the college form, CAT. As an example, Form 1C of the College Ability Test was given to 604 freshmen in 15 colleges, which are listed in Appendix E of the Manual. No information is given as to how these 604 students were selected from among the 1,494 freshmen reported as attending these colleges. A stratified random sample was drawn proportional to the size of each of the 15 colleges, and 370 cases were selected from the 604 for reliability analysis. Comparisons of the total ($N = 604$) population with the sample are made in terms of mean and standard deviation for the verbal, quantitative, survey, and total scores. Each of these scores and four other subtest scores are reported in terms of their Kuder-Richardson reliabilities and their standard errors of measurement. Some question might be raised as to the selection of colleges. One would like to assume they were chosen for their representativeness rather than their availability, but the absence of well-known, large institutions would raise some question as to the appropriateness of these data for general colleges everywhere. Average enrollment of these 15 colleges was 463, and seven of the 15 were described as sectarian in their stated aims and affiliations.²¹ Further reliability studies would greatly strengthen this section of the data.

TECHNICAL RECOMMENDATION

D3.2 The reliability sample should be described in terms of any selective factors related to the variable being measured, usually in-

²¹ Mary Irwin, *American Colleges and Universities*. Washington, D.C.; American Council on Education, 1956.

cluding age, sex, and educational level. Number of cases of each type should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Educational level is the only factor identified in the Manual or its appendixes. While sex and age data are probably not vital in tests of this sort, a *more complete breakdown of the reliability analysis group to portray their identifying characteristics* would be desirable.

TECHNICAL RECOMMENDATION

D3.3 Appropriate measures of central tendency and variability of the test scores of the reliability sample should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This has been done faithfully as indicated in several of the sections above. The actual means of the analysis sample are from 2 to 4.2 raw score points higher than the population from which they are drawn, but since the SD's are approximately the same it can be assumed that this mean difference is within the limits of chance and therefore not statistically significant.

EQUIVALENCE OF FORMS

TECHNICAL RECOMMENDATION

D4. If two forms of a test are made available, with both forms intended for possible use with the same subjects, the correlation between forms and information as to the equivalence of scores on the two forms should be reported. If the necessary evidence is not provided, the manual should warn the reader against assuming comparability. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The coefficient of internal consistency is the type of reliability measure chosen by the authors of this test to describe the reliability of the tests in the SCAT series. The only actual coefficients of reliability reported are of this kind.

TECHNICAL RECOMMENDATION

D5.1 When a test consists of separately scored parts or sections, the correlation between the parts or sections should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

This is presented in a table for each of the two major levels of the SCAT tests, the college CAT and the high school SAT. Inter-correlations are presented for only one form of each of these two levels. *The intercorrelations are rather high (.53 to .62) between the verbal and quantitative scores. This would minimize the use of these subtests for diagnostic or predictive purposes. This point is reinforced in the Manual several times.*

TECHNICAL RECOMMENDATION

D5.11 If the manual reports the correlation between a subtest and a total score, it should point out that part of this correlation is an artifact. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

No discussion of the data presented in the table of intercorrelations is given. *The reader is left to draw his own conclusions. The intercorrelations between the subtests and the total score are high,*

.80 to .89, which are as high as coefficients often reported for reliability estimates in many other tests on the market. This should again warn the counselor that the most useful score for this measure is the total score.

TECHNICAL RECOMMENDATION

D6. Coefficients of internal consistency should be determined by the split-half method or methods of the Kuder-Richardson type, if these can properly be used on the data under examination. Any other measure of internal consistency which the author wishes to report in addition should be carefully explained. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Kuder-Richardson (Formula 20) was used in computing each reported coefficient of internal consistency. The basic assumptions in the use of this specialized formula must be met and clearly stated or there are apt to be some distortions in the values obtained. Of the needed qualifications for use of this formula, only one, the effect of speeded tests upon it, has been adequately stated in the Manual. (See D6.1 below.) The effect of heterogeneity of item content is to yield higher reliability estimates than might be gained otherwise. No mention is made of this possibility in the Manual.

TECHNICAL RECOMMENDATION

D6.1 For time-limit tests, split-half or analysis of variance coefficients should never be reported unless: (a) the manual also reports evidence that speed of work has negligible influence on scores; or (b) the coefficient is based on the correlation between parts administered under separate time limits. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual of the SCAT provides a detailed discussion of the

greater desirability of power over speeded tests in the measurement of educational ability. A table is presented that itemizes the percents of students in the norm analysis groups who completed 100 percent and 75 percent of each of the items as well as an actual item count of the number of items in each subtest reached by 80 percent of the sample. All these figures are high enough to suggest that power rather than speed is essentially being tapped in the SCAT series. As might be expected, slightly more stress is placed on speed in the arithmetic subtests than in the verbal tests.

STABILITY

TECHNICAL RECOMMENDATION

D7. The manual should indicate what degree of stability of scores may be expected if a test is repeated after time has elapsed. If such evidence is not presented, the absence of information regarding stability should be noted. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual clearly states that while no stability data are available at time of publication, studies are under way that will be reported later. This information will be watched for with a good deal of anticipation since the stated purposes of the SCAT series is the continuity of measurement over a long range of years. If this effort is successful, the Educational Testing Service will have filled in a real gap in the field of testing. They will also have provided a real tool for the counselor who is, above anything else, interested in growth and prediction of future success.

E. ADMINISTRATION AND SCORING

TECHNICAL RECOMMENDATION

E1. The directions for administration should be presented with suf-

ficient clarity that the test user can duplicate the administrative conditions under which the norms and data on reliability and validity were obtained. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

A more than average effort is made by the authors of the SCAT series to present the directions for administration in a clear and easily followed manner. A preliminary kind of check list is presented to alert the examiner to have the necessary forms and materials on hand. Some attention is also paid to motivation for test-taking, and care in making proper physical arrangements beforehand. Specific time scheduling is suggested with several alternate plans offered, but no data are given about the effects on scores by the alterations. Attention is given to some of the everyday problems in test administration that are forever plaguing the novice, such as what to do when testees ask numerous specific questions. Proper stress is placed upon the need for uniform and standardized presentation of the test directions and materials.

The actual directions that are to be read are printed in red ink, alternated with cautions to the test administrator printed in the usual black. The red ink alternating with black is also used in the test booklet itself in an effort to increase clarity of procedure.

F. SCALES AND NORMS

TECHNICAL RECOMMENDATION

F1. Scales used for reporting scores should be such as to increase the likelihood of accurate interpretation and emphasis by test interpreter and subject. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT series reports its scores as normalized scaled scores

equated by use of a highly specialized technique, "Lord's maximum likelihood method." The reader of the Manual is left a little in the dark as to just how this equating process is done. The number and characteristics of students used to arrive at this score expression are not described. If the reader accepts this scaled score and its derivation he will find that the Manual claims for it: (1) this scale is not like most normative scales, but is a "test-defined scale" with an unique scale for each ability sample; (2) a particular scale describes the same ability regardless of form or level used, thus allowing comparisons between these forms and levels over a period of time; and (3) the score has no interpretative value in itself, but must be understood in reference to a table of norms. Should all these prove out over the years, the SCAT tests will have made an unique contribution in their scoring system.

TECHNICAL RECOMMENDATION

F4. Local norms are more important for many uses of tests than published norms. In such cases the manual should suggest appropriate emphasis on local norms. VERY DESIRABLE.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Very adequate emphasis is placed upon the desirability of establishing local norms. The school counselor who uses the SCAT should develop local class, school, and system-wide norms. Score distribution forms that may be used to facilitate this calculation come with the tests. Lengthy discussion of the need for and the use of local norms for counseling the individual student is presented in the Manual. Warning is given that "test scores and class averages should NEVER be used as administrative or supervisory 'clubs' to be held over the heads of teachers." Teachers and counselors are given several suggestions as to how best to utilize local normative information in conjunction with the published norms in an attempt to increase the usefulness of the scores.

TECHNICAL RECOMMENDATION

F5. Except where primary use of a test is to compare individuals with their own local group, norms should be published at the time of release of the test for operational use. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

Tables of norms are presented in Appendix B of the SCAT Manual. These are grade norms, and range from Grade 10 to Grade 14. Separate norm tables are given for each subscore and the total score.

TECHNICAL RECOMMENDATION

F7. Norms should refer to defined and clearly described populations. These populations should be the groups to whom the users of the test will ordinarily wish to compare the persons tested. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The Manual of the SCAT series describes its normative groups by naming the schools and colleges from which the scores used in norming were derived. In addition they give the size of the group, state the mean scores and the standard deviations, and identify each school by its geographic location. In this manner 35 secondary schools and 15 colleges were identified as being the coöperating schools in the collection of the normative data.

Only when the counselor begins to inspect these data a bit more carefully does he realize that one weakness of the SCAT at the time of this evaluation lies in its normative data. While recognition needs be given to the fact that the test series is a new one, and that the Manual does state that more data will be forthcoming in the future, both the size and the composition of the normative group

bear some close inspection. A full discussion of the basis of sampling for the high schools involved in the *normative sampling* is given in the Manual. This includes a discussion of the criteria for eligibility established in the selection of the school, as well as a lengthy table that affords the test user a breakdown of the normative sample by region and size of community, by Grades 10, 11, and 12 for the high school level of the SAT (high school) test. The comparative population figures were taken from the biennial report of the U.S. Office of Education. Although size of the normative sample might be increased at least tenfold, the high school norming procedures seem generally adequate.

Much less information is given for the college norms. Not only are the normative figures much smaller in size, but one might suspect some bias in the sampling when the characteristics of the 15 colleges used in the sampling are inspected. As nearly as can be determined, more than one third of the college students are from parochial enrollments in colleges limited to women. At least five of the colleges are so small that they do not appear in standard reference works on college characteristics. None of the colleges has an enrollment of more than 805, and the average size is only 436. One might suspect some differences in characteristics from a broader and more representative sampling of American collegiate youth than those presented in the norm tables. If no such differences are to be expected, it would seem to be the responsibility of the test publishers to make that point clear.

TECHNICAL RECOMMENDATION

F7.1 The manual should report the method of sampling within the population, and should discuss any probable bias within the sample.
ESSENTIAL.

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

According to the Manual of the SCAT series, somewhat different

sampling procedures were followed for the *School Ability Test* than for the *College Ability Test*. In the high school sample the population was defined as a random sample, stratified as to school size drawn from schools in a nation-wide survey of secondary school characteristics. Over 1,800 schools were contacted and approximately 50 percent replied to queries about their characteristics. From this last group, 35 schools were selected on the basis of a combination of geographic representativeness and school size. Just how many of the 900 schools that replied to the queries about their school were willing to participate in the normative study is not clear, nor is the reader given any information as to why only 35 schools were used in the final norms.

The college test population was chosen so as to obtain colleges as nearly as possible like those used in the normative group of the *ACE Psychological Examination for College Freshmen*. No data are presented about how many schools were considered before the final 15 colleges were selected. The college norming sample is compared with the ACE norming sample, but it is done by percentage comparisons of colleges. With a total group of only 15 colleges, a percentage is apt to be somewhat misleading! A reported 50 percent would really be less than eight colleges.

TECHNICAL RECOMMENDATION

F7.2 The number of cases on which the norms are based should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

The SCAT Manual does this reporting, stating that the total normative group consists of 11,829 cases distributed as follows:

Grade 10	3,748	College freshmen	1,494
Grade 11	3,038	College sophomores	953
Grade 12	2,596		

This information is again reported at the bottom of each norm table.

TECHNICAL RECOMMENDATION

F7.3 The manual should report whether scores differ for groups differing in age, sex, amount of training, and other equally important variables. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

No mention is made of sex norms, age norms, or any category other than grade. While this is common practice in tests of this kind, a statement to the effect that expected group differences were small would be reassuring to the test user. Of particular importance to the college counselor would be a clear statement about the admission policies of the colleges used in the normative study and the degree of selectivity they practiced in acquiring their students. Only one of the colleges in the normative group is a state-supported, *coeducational institution of the kind where often the only entrance requirement is high school graduation*. If highly selective entrance requirements were to hold in the other colleges reported in the norms, some questions could be raised about the appropriateness of the norm tables for use with students in colleges with different admission policies.

TECHNICAL RECOMMENDATION

F7.7 *The conditions under which normative data were obtained* should be reported. The conditions of testing, including the purpose of the subjects in taking the test, should be reported. **ESSENTIAL.**

MANUAL OF THE SCHOOL AND COLLEGE ABILITY TESTS

No mention is made of this point in the Manual of the SCAT. The argument is sometimes made that error factors introduced by

such influences would tend to cancel themselves out in the long run. Wide experience with high school youth by the authors would seem to indicate the importance of knowing something about the conditions of the testing situation. The Manual, for instance, recommends that the tests be administered in a single session of 100 minutes. It suggests also, however, that the test may be administered in two separate sessions, one in the morning and the other in the afternoon of the same day. It would be interesting to know whether one plan or the other was used in all the schools participating in the norming program and, if not, whether the two approaches produced different results.

SUMMARY

In this chapter, the reader has been introduced to the Technical Recommendations for Psychological Tests and Diagnostic Techniques, and the importance of this document has been discussed. The document represents an important step toward the future improvement of tests. Its influence can already be noted in the extent to which one publisher has attempted to follow the recommendations in the design, contents, and technical data of one of its new tests. While the SCAT is not a perfect instrument and does not meet all the pertinent standards, many of its weaknesses and limitations are acknowledged in the test Manual. This, it seems certain, is due to publisher's attempts to follow the standards presented in the Technical Recommendations. The potential user of the test is provided, in general, with the type of information upon which to base a decision as to whether or not it meets his needs and, if selected for use, the limits placed on such use. It is hoped that other publishers will similarly model their future tests around the recommendation as they exist at present and as they may be modified in the future. While the Technical Recommendations may do much in improving the content of test manuals and tests

themselves, the counselor will need to continue his critical appraisal of tests, for the ultimate responsibility for their use is still his.

DISCUSSION QUESTIONS AND EXERCISES

1. It is held by some authorities on measurement that "standards" inhibit experimentation and innovation. Do you agree or disagree with this position? What arguments can be made for each side?
2. From reading previous chapters of this book and reviewing a copy of the Technical Recommendations, are there any aspects of testing that you believe are not adequately covered by the standards? What additions would you recommend?
3. After reviewing a copy of the Technical Recommendations, evaluate those standards categorized as *very desirable* and *desirable*. Would you, in principle, have placed them in these categories? What loss, if any, is suffered if a test manual omits discussion of these standards?
4. Using a copy of the Technical Recommendations as a guide, analyze the manual of a test published before 1954. To what degree is the manual deficient in terms of the recommendations? Is the treatment of some areas better than others? How would you account for the difference in treatment?

REFERENCES

- American Educational Research Association. *Technical Recommendations for Achievement Tests*. Washington, D.C., American Educational Research Association, 1955.
- American Psychological Association. "Technical Recommendations for Psychological Tests and Diagnostic Techniques." Supplement to *Psychological Bulletin*, March, 1954, 51. Washington, D.C., American Psychological Association, 1954.
- Educational Testing Service. *Examiner's Manual, School and College Ability Tests*. Princeton, N.J. Coöperative Test Division, Educational Testing Service, 1955.

CHAPTER VI

Recording and Reporting Test Scores

Some sort of cumulative record of a student's progress has become almost a *sine qua non* in educational institutions and an essential part of that record is a section devoted to test scores. The test section can become a sterile, uninformative, even confusing addition to the records or it can contain vital, revealing items of significance for use in the counseling process. If the tests have been carefully selected, they will be made most useful if the scores derived from them are recorded clearly, cumulated effectively, and made available for ready interpretation by those who are to use them. This chapter will be devoted to a discussion of how that can be accomplished.

At the time a test is given it is not always possible to predict at what later time and by whom the scores will be used. Various persons may find it necessary to look at a pattern of test scores achieved by a student over a long period of time. Scores obtained early in a student's career must be available in such form that comparisons of them with later scores can be done readily by counselors, parents, teachers and other school personnel, employers actual or potential, admissions officers of higher education institutions and, occasionally, persons who have such special assignments as

selecting among candidates for scholarships. Serving the purposes of persons who are likely to have widely differing knowledge about tests requires that several difficult decisions be made about the recording and interpretation of test scores.

METHODS OF RECORDING

Test results are commonly recorded in tabular or graphic form but sometimes they are presented in paragraphs. An informal poll of the preferences of school personnel for these three methods was conducted by the Records and Reports Committee during the *Eight Year Study of the Progressive Education Association*.¹ They were almost unanimous in their selection of the tabular method. At a later time they were given the choice between the two forms shown in Figure 2 for recording test results and again the vote was overwhelmingly in favor of the first.

It seems, then, that it will be advisable to record the test scores in tabular form but the question of what will be reported in the tables must be considered. The full name of the test (or an abbreviation that cannot be misinterpreted) and the form of the test must, of course, be recorded. The date of administration will be important but it will not usually be necessary to indicate more than the month or year if the subject's school grade is on the cumulative record. The common practice of recording raw scores is unnecessary if the derived score is given and description of the norm group is clearly stated, but when local norms are used it may be necessary to report the raw score. Thus, in the example in Figure 2, it might be desirable to eliminate the raw score column and substitute one in which the norm basis could be noted, or to add another column so that both raw scores and norm basis could be indicated.

In making decisions about which form of derived score to use, several factors must be considered. The counselor must be concerned primarily, however, with the problem of interpretability to

¹ E. R. Smith and R. W. Tyler, *Appraising and Recording Student Progress*, New York: Harper & Bros., 1942.

[illegible]

counselees and the minimizing of possible errors of interpretation by those persons who may have occasion to use the data.

The Psychological Corporation has provided an excellent diagram (see Fig. 3) of methods of expressing test scores. It presents

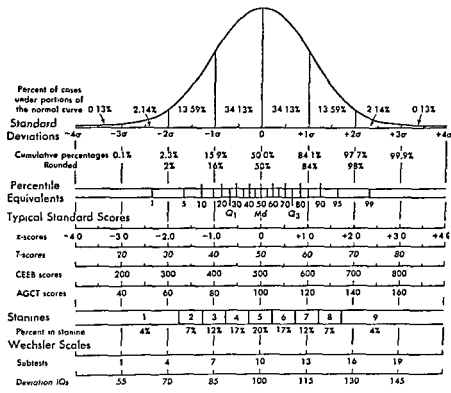


FIGURE 3.

all the common methods of reporting scores and adds others such as the Wechsler Scale and Deviation IQ methods, which may be used occasionally.

The difficulties in interpretability of a set of test scores for the several persons who are to use them will not be lessened if several different methods of recording them are used. It seems desirable to choose one form and retain it throughout a student's record. When an occasional deviation from that form, as in the case of the

Wechsler Scales, is used, it will be desirable to supplement the record by an explanatory paragraph.

Recording of derived scores in terms of percentiles seems likely to be most meaningful to counselees and other persons who are not familiar with esoteric terms used in measurement. Percentiles have the disadvantage, as shown in Figure 3, that they are not equal points on a scale and therefore a score that is at the 95th is farther away from the 85th percentile than a score at the 55th is from one at the 45th percentile. A further disadvantage is that percentiles may be confused with percentages. Both of these difficulties and sources of confusion are easier to clear up to a counselee than are the problems that arise when attempts are made to explain what is involved in any of the forms of standard scores.

It has been found that a vast majority of a sample of test users prefer a percentile equivalent for every possible raw score.² Using such percentile tables, it is possible to show a counselee where he stands in relation to some defined group and to show him that the differences at the ends of the scale are greater than those at the middle.³ The tables can also be used to show the differences between percentile and percentage.

When the test data have been recorded by percentiles in tabular form on the cumulative record (and there seems good reason to believe that even the intelligence test scores would be recorded in percentiles) the task of recording is not complete. The tester should record any peculiar conditions in the test situation, his observations of students' reactions, or any facts about a student's general characteristics that might have affected his scores. Thus, if it is known that a student is meticulous in checking every bit of his work and is thus likely to be handicapped in speed tests, that fact should be recorded. If a student is known to be exceptionally

² *Test Service Bulletin*, No. 48, New York: The Psychological Corporation, January, 1955.

³ Two of the authors have done so for high school students and have indicated their results. See J. W. M. Rothney, "Interpreting Test Scores to Counselees," *Occupations*, February, 1952, pp. 320-322.

nervous about all types of objective tests, a note to that effect should be added. If still another is known to be particularly disturbed by certain kinds of tests, such as those that require the use of numbers, a notation about the circumstance should be made in the section devoted to test interpretation. The following samples abstracted from some student cumulative folders illustrate such comments.

This boy finds all tests so easy that he usually hurries through them, and spends the rest of the testing period watching his fellow students.

Jane is so flustered at all objective testing sessions that she cannot seem to get started until the other students have finished a good part of the test.

Jim goes through a test quickly, doing the items that he knows he can do first. He then goes back to the more difficult items.

Dian said that she just could not do arithmetic and that she would never, even if other phases of it were excellent, take a job where she would be required to make change for customers.

Clark can never shed his scorn for things academic long enough to put forth maximum effort on tests so that all his scores are questionable.

Larry's scores always seem lower than one would expect from him. His effort was not always maximum even when test situations are carefully proctored.

As indicated in Chapter VII, many factors within and without the individual may influence his performance on a particular test. If those factors are of such strong influence that the test score must be seriously questioned, it will often be better to omit it from the test record completely and to note on the record the reasons for doing so.

PROFILES

When a student's test scores have been recorded graphically, the result is sometimes described as a profile or psychograph. The object of the test profile is to set forth test results in the simplest and clearest form for the user of test results. It may be a line diagram that indicates the relative position of divergent standing on various tests as well as the overall picture of his performances. It may be a curve joining the successive points on a graph representing individual status in each of several traits. In such cases it is a method of recording status of an individual in each of several traits so that the geometric pattern created may be meaningful. These profiles are often used uncritically despite occasional reminders by measurement experts that there are many pitfalls in their use. Such factors as independence of the variables, inequality in scales, reliability of subtest scores, and the relative importance of the variables must be considered if profiles are to be used.

Since profiles are merely graphic representations of what may be fallible test scores, they must be interpreted in terms of the varying reliabilities of the tests and in terms of the intercorrelations among the tests represented. If test results are to be used in profile form, it is essential that the scores of all the tests be reduced to a common origin and a common unit of measurement. Psychologists and educators have rarely been able to devise compensating scales in which units on one scale represent equal amounts of the properties being measured by another.

At this point the student should look carefully and critically at the profiles obtained from such tests as the California Test of Mental Maturity, The Stanford Achievement Test, and the Differential Aptitude Tests, and at the profiles obtained from such questionnaires as the Kuder Preference Record or the Minnesota Multiphasic Inventory. He should be sure to examine the data in the test manual about the standardization populations, the inter-

correlation of subtest scores, the reliability of the subtests, and the evidence (if he can find any) that the subsections are of somewhat comparable validity and importance. If the subtests were standardized on populations differing in quality or number, if the inter-correlation coefficients are high, if the reliabilities vary greatly, and if the evidence of comparable validity and importance is incomplete, the value of the profile must be questioned. If the answers to the questions, What are the important factors to be measured? How can really comparable measurements be made of those factors? are not available, it will probably be desirable to omit the use of profiles.

Much of what has been said about profiles applies equally well, of course, to a column of test scores. The same precautions in interpretation are necessary when scores are in the usual tabular form, but the profile causes more difficulty because sharp peaks and hollows indicated by the (almost meaningless) lines of the profile are likely to stand out and seem to mean more to the counselee or other untrained observer than they should. The elation at the sight of a peak on a profile or the despondency at a sharp decline of a curve is likely to leave lasting impressions on a counselee or his parents. The interpretation that they draw from such a profile may be out of all proportion to its importance despite the precautions advanced by a counselor.

WRITTEN REPORTS

As suggested previously, scores from a test without some verbal interpretation are likely to be misinterpreted and those previously posted on a record may be misleading. In the case of tests that are administered individually, the recording of a score without any comment may result in the loss of many of the observational data that may be as important as the test score. To avoid such loss, a paragraph statement should be written. The following statements are samples of the kind that may be used.

John's IQ score of 128 places him among the upper ten percent of the population in the type of performance measured by the tests. He tried very hard and his anxiety to succeed was shown by his repeated questions concerning the correctness of his answers. He did exceedingly well on all the items that required use of figures and nonverbal material but performed only at average levels for his age on verbal material. He was an alert, interesting boy who spoke very freely about himself and his many activities. He is left-handed. Frequent eye-blinking suggests the need of ocular care or investigation concerning situations that provide too great emotional stimulation.

Jane's IQ score places her in the lower five percent of the population in the type of performance measured by this test. She tried hard and failed often but did not seem to care whether she succeeded in passing tests. She failed to respond to praise and encouragement and seemed to be very glad to finish with the testing. Her attention wandered frequently and many responses seemed to be unrelated to the tasks she was required to do. She asked for many repetitions of directions and there is the possibility of a hearing difficulty that ought to be investigated. She has a decided squint. We suggest that she be given a thorough investigation to determine whether classification of mental deficiency can be made.

Paul's IQ score of 135 places him in the upper 2 percent of the population in the type of performance measured by this test. He did many tests four years beyond what is expected of a boy of his age and he did all kinds of tests on this scale with equal facility. He took the test as a challenging game and we feel very sure that this record describes his best performance. The behavior problems that this boy exhibits in school might well be due to the fact that fifth grade work is not challenging enough for him. We would like to give him several other tests and to examine other information about him to determine the advisability of promoting him to the junior high school.

Two difficulties with the use of paragraph reports are the time

that is required to write them and their subjective nature. If it has been worth taking an hour of a student and counselor's time to have the test administered, it is probably worth the extra few minutes it takes to write the report. The statements in the report must, however, be considered only as those of a particular observer in a specific situation. Generalizations about a counselee that are based solely on the paragraph reports must not be made. If they corroborate other evidence that has been obtained previously, that corroboration may make the evidence more useful. If they do not, further investigation will be needed to determine whether or not the subject's performance in test situations is really significant enough evidence of variability from usual performances to warrant consideration in planning with him. In either case the results may be useful in his counseling.

CUMULATIVE RECORDING OF SCORES

Some of the problems of recording and making available the test data of an individual would be serious enough if they referred to the scores obtained in only one year. The problems become vastly more complicated when longitudinal data on test performances are obtained. The interpretation of a set of test scores currently obtained is fraught with difficulties, but they seem to be minor compared to those that one meets when attempts are made to interpret a series of scores obtained over a period of years. It is difficult to determine whether a counselee has grown in the characteristics that the tester has attempted to measure because it is almost impossible to get comparable units of measurement.

In the area of personality assessment, the problem of getting meaningful units of measurement seems impossible of solution at the present time. There is nothing in that area that even approximates the metrics of difficulty that can be used with fair success in the fields of achievement and intelligence testing. In the latter areas it is assumed that the greater number of questions one can do, the

better the score will be; but that assumption cannot be justified in personality measurement. Is it better, for example, to be 100 per cent responsible, extroverted, dominant, creative, free from use of defense mechanisms than it is to be partially so? Is the golden mean better than the extremes? Since in this area the best is not *certainly* the most, those who would assess personality cannot use the metrics of frequency and difficulty that the achievement tester uses. At this time no satisfactory substitute has yet been found.

In the fields of achievement and aptitude testing, a common practice is to allot certain values to correct answers to questions on a test (usually one point), to count them, and to conclude that the highest total is the best score. Still another method is to give test items to different age or grade groups and, by comparing their scores, set up age or grade norms. It is then assumed that the highest scores made by the older subjects or higher grade groups are evidence of mental development or growth in achievement. In the process, assumptions of content and comparability of metrics are made. If such assumptions were always tenable, the interpretation of cumulative test records of an individual would be less difficult than it now is. Test-makers have seldom given test users enough data to assure them that scores obtained on the same subjects over a period of years are comparable.* They cannot be sure that equal increments in test scores are associated with equal increments in the characteristics that are presumed to be measured by the test. So far, the best that test-makers can do is to compare a subject's performance to those of a defined group in terms of the distribution of scores of the group and to repeat that comparison with a later group. The units of measurement employed in the comparisons are percentiles, age, grade scores, or one of the several forms of standard scores variously labeled as T-scores, scaled scores, stanines, and z-scores. They are illustrated in Figure 3. No one of them avoids the difficulty that they do not make equal units of

* The efforts of the Coöperative Test Division of the Educational Testing Service in this direction have been particularly praiseworthy.

measurement except by definition. Tiedman points out, however, that this procedure is not unique to educational measurement in these words: "However, description is possible through use of defined units of measurement. Measurement of such concepts as temperature and speed has proceeded by means of defined units. The practice of defining units of measurement may be defended as valid as long as the results of experiments involving the units as defined prove useful and understandable; that is, as long as the results are consistent with expectations for them, were equal increments of normalized scores associated in reality with equal increments of the characteristics measured."⁵

After consideration of the values and limitations of several methods used in attempts to get equal units of measurement Tiedeman concludes that in terms of comparability the methods may be ranked as follows: (1) K-scores, (2) scaled scores, (3) T-scores, (4) age and grade scores, and (5) percentile ranks.⁶

Tiedeman has ranked percentiles as lowest in terms of comparability. It will be noted that the writers have placed them first with respect to their interpretability to counselees and other nontechnically trained school personnel or parents. If the claim of highest ranking for percentiles in interpretability is valid and Tiedeman's claim of lowest ranking for them in comparability (and hence as poor measure of growth) are both justified, the counselor who wants to put scores on cumulative records must find himself in a dilemma. If he reports scores in terms of percentiles for better current interpretation to his counselees, he will be less sure of his data when he is called upon to answer questions about whether the student has been developing well, normally, or poorly. And it has been suggested frequently that longitudinal measures of development will be better predictors of future development than a cross-sectional picture obtained from one set of scores.

⁵D. V. Tiedman, "Has He Grown?" *Test Science Notebook*, No. 12, Yonkers, N.Y.: World Book Co.

⁶See references at the end of the chapter for descriptions of these methods.

Unfortunately, no crucial experiments have been carried out to help the counselor who recognizes the dilemma. He will be forced to rely on his own judgment. He may recognize that all methods have limitations and may still prefer to utilize percentiles in his interpretations to lay persons. He will also be aware that his judgments about growth must be tempered by his knowledge of the limitations of the technique. If he wishes to make studies in growth he will work out the more technical units of measurement from his raw scores, but the entering of such scores on the cumulative record is likely to result in more confusion than clarity to most users of the record.

PREDICTION AND TRANSFER

Much of what has been said previously in this chapter applies to the handling of test scores when they are to be sent on to schools and colleges or to potential employers. Except in rare cases, the form in which data are to be recorded when the student transfers is not specified and the counselor has a choice of methods by which he can meet his responsibility of making the scores meaningful.

Unless the counselor has evidence obtained from expectancy tables or has, from some other source, become thoroughly familiar with the situation into which the counselee proposes to enter, he should not make predictions about probable success in it. It is the responsibility of an admissions officer or employer to make the decision as to whether an applicant will be accepted. If he accepts him, it is implied that he expects (predicts) successful work from the student or employee. If his prediction is wrong and the subject fails, he cannot then place the responsibility for that failure on the counselor who could not possibly know as well as the admissions officer or employer all the circumstances that the subject may find.

The counselor may, as a guide to future action, make predictions of his own concerning the success of his counselees who go on to particular employment or training situations. From follow-up data

he can check his predictions and use the results to help future counselees who plan to enter such situations. He would very probably meet serious difficulties, however, if he were to predict success of counselees in situations with which he was not thoroughly familiar. Only occasionally will he find an institute of higher education that has worked out tables about performances of entrants from which a candidate for admission can estimate his chances for success. It will be the responsibility of the counselor to inform the student about such tables when they are available.

The figures that follow were derived from two sets of data obtained about each member of a large freshman class of one university and a study of their achievement during their first semester. They indicate the percentages of freshmen who achieved less than a 1.0 grade point average (which meant that they would be dropped from the institution), of those who achieved better than a 1.0 average, and those who gained greater than a 2.0 average (better than B grades) if their test scores lay within each of the four quarter of two sets of data.

The counselor may discuss this table and the test scores with a student who is applying for admission to this institution but he should not predict that the potential student would achieve at the level of any of the grade-point categories. He should be certain to make clear what odds probabilities, or chances of success were based upon the performances of a group who preceded him. They do not *certainly*, despite the headings in the table, indicate the chances of success of any particular student. The potential student *whose scores lie in the top quarters cannot assume that he "has it made."* Nor can the counselee who scores in the lower quarters assume that failure is certain. Both can probably profit from a discussion of what others have done and what they can do about decisions to apply for admission and the level of work they must do if they are accepted.

Unfortunately, there are not many educational institutions or

TABLE 17. Possible Achievement Based on First Semester Study of Freshmen

<i>Code</i>	If a student scores in the following quartiles on the American Council on Education Psychological Examination and his high school centile rank is	His chances of obtaining less than a 1.0 average for his 1st semester work are <i>approximately</i>	His chances of obtaining a 1.0 average or <i>better</i> for his 1st semester work are <i>approximately</i>	His chances of obtaining a 2.0 average or <i>better</i> for his 1st semester work are <i>approximately</i>
1-1	1st on both	5 out of 100	95 out of 100	60 out of 100
1-2	1st on one, 2d on the other	15 out of 100	85 out of 100	35 out of 100
1-3	1st on one, 3d on the other	20 out of 100	80 out of 100	20 out of 100
1-4	1st on one, 4th on the other	30 out of 100	70 out of 100	10 out of 100
2-2	2d on both	40 out of 100	60 out of 100	10 out of 100
2-3	2d on one, 3d on the other	55 out of 100	45 out of 100	8 out of 100
2-4	2d on one, 4th on the other	65 out of 100	35 out of 100	2 out of 100
3-3	3d on both	65 out of 100	35 out of 100	2 out of 100
3-4	3d on one, 4th on the other	85 out of 100	15 out of 100	1 out of 100
4-4	4th on both	90 out of 100	10 out of 100	Less than 1 out of 100

employers who have data comparable to those given above. Without them prediction of success is a hazardous procedure. Even when some unavoidable circumstances (such as the demands of some colleges that a prediction of achievement be made for all applicants) force the counselor to record a prediction on an application blank, he should indicate very clearly that, even with the best tests currently available, his prediction must always be qualified.

Prediction of counselee's performances is a problem about which counselors must be acutely sensitive. Test builders and other measurement persons tend to claim higher prediction performances than they can justify, ignore the issue completely, or dismiss some failures by indicating that if only a few "false positives" appear the attempts to predict have been justified. Even one false positive, in the person of an influential individual's son for whom success in college was predicted and who failed to make the grade, may result in irreparable damage to a counseling program. The counselor will do well to limit himself to descriptions of his subject's performance in clearest possible form, to interpretations of his data in terms of probabilities when he has enough evidence to do so, and to leave predictions to those who must assume responsibility for the subject in the employment or training opportunity that he enters.

JOINT USE OF TESTS

The counselor is not the only person in an educational institution who is concerned about tests. Administrators want some data on the general status of the students, and teachers are interested in the level at which their pupils perform. It has been suggested that if a well-selected battery of achievement tests is given for administrative and teaching purposes they may serve equally well for counseling. *The counselor may be interested in getting evidence of a student's current achievement on mathematics tests in which the items cover the rules and symbolic systems that have been taught.*

For the counselor's purposes it may be more useful to have the records of performances on an achievement test that a teacher selects because it covers the area she has taught than to have scores from so-called numerical aptitude tests that sample materials with which the student has had no experience. The test that the teacher of typewriting selects may give the counselor a better measure of performance in an important part of stenographic work than a so-called *stenographic aptitude test*. Reading achievement test scores may be as useful to the counselor as to the teacher or remedial reading specialist. In view of these common interests in test scores, it may be possible to use the same test battery for administrative, teaching, and counseling purposes.

There are certain limitations that must be noted, however, if the tests are to be used by the three groups. The use of test batteries that are administered to all the students tends to discourage the practice of tailoring the testing program to fit the needs of particular counselees. Since the tests used for teaching purposes are likely to be subject matter tests, some nonacademic students may be discouraged and may not do well. Tests that may have excellent content validity and therefore cover a given course area may have little predictive validity. To compensate for these limitations, however, an achievement battery has the advantages of assuring that the subject has been tested over areas covered in instruction. The joint use of a battery of achievement tests may save time and money and it may result in bringing the counseling program closer to the faculty. And it may tend to reduce the artificial gap that has seemed to separate achievement and aptitude testing.

Although the counselor participates in the selection and use of an all-school achievement testing program, he will probably want to supplement the tests with some that are designed particularly for use in counseling. He will want to use selected tests to be given to particular counselees at particular times. When the tests are given for any of these purposes, however, it is essential that the

suggestions for recording and interpreting test scores previously described in this chapter be given full consideration.

SUMMARY

This chapter has been concerned with the problems encountered *in recording and reporting of test scores in a manner that provides for their best use by counselors and others.* It was suggested that test users prefer tabular to other methods of recording and that counselors will find the percentile procedure most useful in test interpretation, although more precise methods may be most useful *for measurement of growth, and in research.* Inadequacies of profiles were described and the need for written description of subjects' performances in test situations noted. Problems of prediction in the transfer of counselees from one educational level to another were described and the advantages and disadvantages of the joint use of tests was considered.

DISCUSSION QUESTIONS AND EXERCISES

1. In the following table the numbers and percentages of students who scored in each of nine categories (stanines) in a state-wide

TABLE 18. Number of students in each test category

Stanine	9	8	7	6	5	4	3	2	1	Total
Satisfactory college work	43	57	79	43	28	9	1	0	0	240
Unsatisfactory college work	2	11	16	39	39	25	8	1	2	143
Totals	45	68	75	82	67	34	9	1	2	383

testing program administered when they were in high school are given. (The top category is nine, the lowest is one.) The success of the students in their first year at X college is also given.

- a. A counselee who has taken the state-wide tests and scored in

- the top (9) category wants to know what his chances of doing satisfactory work at X college are. What could you tell him?
- b. A counselee who scored at the (3) level asks you the same question. What would your answer be?
 - c. What are the approximate odds that a student who has scored at the (4) level will fail to do satisfactory work at X college? What odds if he scored at the (5) level? Is it true that there are only two possibilities, succeed or fail, and that odds cannot be stated?
 - d. What could you tell a counselee who had scored at the (5) level about probability of doing satisfactory work at another college?
 - e. The parents of a student who scored at the (1) level insist that their son, your counselee, attend X college. What would you tell them?
2. The following set of figures indicates the percentile position in a university freshman college class of students whose scores fall at specified percentile ranks in a state-wide testing program for high school seniors.

TABLE 19. College Freshman Rank Indicated by High School Percentile

If Percentile Rank in High School Was	Rank in the Freshman Class of X College Would Be
41	10
59	20
68	30
75	40
80	50
86	60
91	70
94	80
97	90

- a. One of your counselees scores at the median for high school students. He says that he wants to go to X college and wants to know if he is smart enough to succeed. What can you tell him on the basis of the data in the table?

- b. Another counselee scores at the 80th percentile in the high school testing program. He has found high school work easy and expects that the work in X college will be a "breeze." What can you tell him?
3. It is said that the students who score in the upper third of their high school classes on mental tests should go to college. What do the figures in the table suggest about such statements?
4. The following quotations were taken from educational autobiographies written by university students in a first course in education. What questions do they raise about the use of tests, inventories, and personality questionnaires in high schools?
- About this time Margie and I decided that we wanted to know what our IQ's were, so we sneaked into the office one night when something was going on at school and looked. Actually, we weren't too afraid of being caught, because we felt we had a right to know.
 - During high school I never thought much about going to college, and thought I would become a secretary upon graduation from high school. I liked my business courses, and thought I would succeed if I went into a business career. On an aptitude test taken during my sophomore year, I ranked high in clerical and secretarial fields. This made me even more positive of the worth of my secretarial career. Always in the back of my mind had lurked the desire to be a teacher, especially of English.
 - I was called into the office one day while in high school to take a social aptitude test. Unfortunately I was "mad at the world" that day. A few days later I was *rushed* into the office and they started asking me questions. But everything got straightened out and me and society were again O.K.
 - My high school was kind of an experimental school. It seems as if I took tests constantly. During my last semester, I had a senior conference with a teacher of my choice. She showed me the results of all my tests and interpreted them for me. The conference was extremely beneficial to

me for it pointed out my strong and weak points, my likes and dislikes. I don't remember what my IQ was but she said I had definite potentialities for a successful college student. My activities and interests were in music, economics, working with money and figures, and dealing with people. The conference didn't sway me towards a different vocational career but instead, it just strengthened the ideas and desires I had already formed. I certainly think these tests should be made available to all seniors in high school.

- c. Through a standard series of tests as to my capabilities and interests I learned in the ninth grade that my interests and the most likely region for success were in teaching of science or doing active dramatic work. Throughout the examination, I listed what I knew that I *should* do rather than what I actually *prefer* to do.

REFERENCES

- Cook, W. W. "What Educational Measurement in the Education of Teachers?" *Journal of Educational Psychology*, April, 1950, pp. 339-347.
- Courtis, S. A. "Personalized Statistics in Education." *School and Society*, May, 1955, 81:170-172.
- Flanagan, John C. *Bulletin Reporting the Basic Principles and Procedures Used in Development of a System of Scaled Scores*. New York: The Coöperative Test Service, 1939, 41 pp.
- Gardner, Eric F. "Value of Norms Based on a New Type of Scale Unit." *Proceedings of the 1948 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1948, 117 pp.
- McCall, W. A. *Measurement*. New York: Macmillan, 1939, 535 pp.
- Rothney, J. W. M. *Appraising and Recording Pupil Progress*. Washington, D.C.: Bulletin No. 7 of the American Educational Research Association, 1955, p. 30.

- Rulon, P. J. "On the Concepts of Growth and Ability." *Harvard Educational Review*, 1947, 17:1-9.
- Smith, E. R., and Tyler, R. W. *Appraising and Recording Student Progress*. New York: Harper, 1942.
- Thurstone, L. L. "A Method of Scaling Psychological and Educational Tests." *Journal of Educational Psychology*, 1925, 16:433-451.
- Traxler, A. E. "Evaluation of Methods of Individual Appraisal in Counseling." *Occupations*, November, 1947, pp. 85-91.

CHAPTER VII

Combining Test Scores with Other Data

In various chapters of this volume it has been shown that the value of a test score for a given counselee must depend on such factors as the validity, reliability, and norms of the test. After a counselor has used an instrument that was carefully selected on the basis of such factors and obtained a score for a counselee, he will find it necessary to supplement the test score with information about past performances. If such information does not contain scores from other tests, the counselor must operate temporarily on the assumption that his counselee's test score is a dependable sample of his usual performances. He can seldom find adequate evidence that it is. As he examines the test performances of a counselee, he may become aware that such specific factors as set, fatigue, indifference, nutrition, motivation, purposes in life, guessing, memory, rapport, as well as more generalized factors of socioeconomic class, previous testing experiences, or attitudes toward self and authority figures may have been operating to a greater or lesser degree.¹ Such factors may be crucial in interpretation of the test performances for a given individual.

¹ Robert L. Thorndike. *Personnel Selection*. New York: John Wiley & Sons, 1949. A fully developed outline of intrapersonal influences on test performance is presented on page 73.

Faced with the realization that any one or any combination of such factors may have prevented his obtaining a dependable test score, the counselor must attempt to determine the relationships among test scores and other data to be used in counseling. To do so he may compare test scores with other evidence about his counselee on the assumption that he will be somewhat consistent in several kinds of performances. Perhaps, in the case of a particular counselee, the current scores are in line with previous academic and test achievements. If they are, the counselor may feel that he has some dependable evidence about his counselee since it has been demonstrated that relatively high consistency of performance can be expected for about one half of a group of high school students when test scores, course grades, and teachers' descriptions of behavior are obtained over a period of several years.² A body of theoretical writing^{3, 4, 5} in the field of personality organization also bears out the contention that a generalized consistency of traits and behavior is common to many persons.

Viewed in this manner, it is necessary in making interpretations of test scores to use longitudinal data that commonly appear on cumulative guidance records of counselees. With such data the counselor may look at his counselee's observed and recorded developmental history rather than at the very brief sampling of performance provided by the thirty minute or even the three-hour test battery. By judicious combination of scores from tests and longitudinal clinical data, he may find some answers to the questions his counselee has raised.

COMBINING TEST AND CLINICAL DATA FOR COUNSELING

High school students frequently ask their counselors, ". . . am

² Robert A. Heimann, "Intra-Individual Consistency of Performance in Relation to the Counseling Process." Unpublished Ph.D. Dissertation, University of Wisconsin, Madison, 1952.

³ Gordon Allport, *Personality*. New York: Henry Holt & Co., 1937.

⁴ Gardner Murphy, *Personality: A Biosocial Approach*. New York: Harper & Bros., 1949.

⁵ Prescott Lecky, *Self-Consistency*. New York: Island Press, 1945.

"I bright enough to succeed in college?" The conscientious counselor must always preface his answer to this kind of question with a cautious modifier, ". . . that depends. . . ." The answer to a seemingly simple and direct question such as the one above must depend upon at least two major types of information. The first is introduced by the test score itself. The second requires consideration of the individual as a person and of his background.

INFORMATION PROVIDED BY TESTS

When trying to help counselees find answers to questions such as the one raised above, counselors commonly give such tests as the Ohio State Psychological,⁶ the Otis Quick-Scoring,⁷ the American Council on Education Psychological Examination for College Freshmen,⁸ or one of the others said to measure scholastic aptitude. Having administered the test, the counselor then looks at the score and compares it to others given in the table of norms. If his counselee's score lies near or above the 75th percentile, he is apt to support the counselee in his plans to go to college. If the score is close to the 25th percentile, he is apt to be less than encouraging over the prospects of the student's success. At this point he should combine the data provided by the test score with information he has obtained about typical performances of his counselee from other sources. In this way he may avoid serious consequences of misinterpretations of a single test score.

PREDICTION FROM TEST SCORES

The counselor will have examined the claims for validity of a

⁶ Herbert A. Toops. *Ohio State Psychological Test*. Chicago: Science Research Associates, 1941.

⁷ Arthur S. Otis. *Otis Quick-Scoring Mental Ability Tests*; New Edition. Yonkers, N.Y.: World Book Co., 1954.

⁸ L. L. Thurstone and Thelma G. Thurstone. *American Council on Education Psychological Examination for College Freshmen*. Princeton N.J.: Educational Testing Service, 1954.

test to determine the extent to which its scores predict college success before he administered it. Most research workers who have used test scores in efforts to predict success in training (or on the job) report coefficients of correlation between test performance and future grades or other criteria of success in the range of $r = .40$ to $.50$. Relationships of such magnitude allow the counselor to improve his group predictions over chance only about 12 to 15 percent.⁹ In some instances he may combine test scores with such other predictors as rank in high school class and obtain multiple correlation coefficients as large as $.60$. Coefficients of this size will increase the accuracy of his average predictions to approximately 20 percent above pure chance.

A prediction that raises an estimate by 20 percent may be considered good at the race track. It leaves something to be desired when the stakes are success or failure in the post high school career of a counselee. When the counselor recognizes that his tests provide estimates that improve upon guesses only to this modest degree, great care must be taken in test score interpretation. He must avoid dogmatic, categorical statements such as: ". . . you should give up the idea of going to college because your score is too low . . .," or ". . . you will never make it with these scores . . .," or ". . . you have the ability to succeed in training for the career of your choice because your score is high on this particular aptitude test. . . ." But he is obligated to inform his counselee of his chances for success upon information gathered from actual research studies. If such data are available his interpretation might be stated generally, ". . . 20 out of 50 students from our school with test scores such as yours have been successful in the training in which you are interested."

⁹ The statistic "k" or coefficient of alienation from which such percentages are computed may be calculated for any degree of relationship. See: J. P. Guilford, *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill Book Co., 1956, pp. 375-379.

EXPECTANCY TABLES

The data from which a statement such as the above can be made are usually reported in an expectancy table. If such tables can be drawn up and used in test score interpretation, much of the mystery surrounding such scores can be cleared up and better communication between counselor and counselee developed. The expectancy table shows relationships between individual test scores and criteria of success more clearly than the general trends indicated by coefficients of correlation, which counselees cannot be expected to understand. It is an effective method of helping teachers and parents to see that the relationship between scores on tests and criteria such as grades is not perfect, and that there is a great deal of overlap in the grades obtained by students who make different test scores. An example of one expectancy table that illustrates the relationship between performance on the Differential Aptitude Tests Sentences section of the Language Usage subtest and course grades in an English class is presented below.

TABLE 20. Relationships Between Grades in English and Differential Aptitude Test Language Usage Scores*

Number of Cases	Number Receiving Each Grade in Class in English					DAT Test Scores	Percent Receiving Each Grade in Class in English					Total Percent
	F	D	C	B	A		F	D	C	B	A	
1					1	80-89					100	100
5				1	4	70-79				20	80	100
22			3	14	5	60-69			14	63	23	100
23			9	8	6	50-59			39	35	26	100
22			3	13	6	40-49		14	59	27		100
16		1	3	9	3	30-39	6	19	56	19		100
8		1	4	3		20-29	13	50	37			100
2			2			10-19		100				100
1			1			0-9		100				100
100		2	13	37	32	16						

SOURCE: Alexander G. Westman, "Expectancy Tables—A Way of Interpreting Test Validity," *Test Service Bulletin*, No. 38. New York: The Psychological Corporation, December, 1949.

* Frequencies appear in the left section of the table and percentages are presented at the right.

For illustration it can be assumed that the counselor using this table is concerned with a *freshman who made a score of 55 on this test*. Of the 23 students who scored between 50 and 59, none received a grade lower than C. He may then tell the freshman that, on the basis of past experience, all students who scored as high as he did had made grades of C or better. This does not mean, of course, that this particular student might not be the exception. But it does give him a probability statement in nontechnical language that he is able to see and understand.

The data in the table illustrate the difficulty in prediction from scores in middle ranges. It may be noted that equal numbers of students with scores between 30 and 39 made B's and D's, although one received an F and slightly over half of the students with this score did achieve average grades of C. Data obtained from sources other than tests may prove helpful in this instance if they are considered by the counselor along with the test score.

ERRORS IN PREDICTION. Before the counselor can use such statistics to help in test interpretation he must demand that certain conditions be met. He must satisfy himself that his counselee does not differ markedly from the students used in reported studies with the tests he is considering in terms of their home background, socioeconomic and cultural level, and educational opportunity. He should test the assumption that the grading practices and the range of student talent in the schools used in the reported validity studies are not greatly different from the schools his counselee has attended. And he should be sure that the reported coefficients are based upon sound sampling and carefully controlled studies.

The reader should keep in mind that all statistical presentations of relationships between test performance and future success are based on the *averages of groups*, and may or may not be appropriate for *particular persons*. Consequently any efforts at prediction based upon the average of the group may be less than perfect for the individual members of it. The error in prediction resulting from

this fact is similar in a general way to the error that accompanies each test score as an estimate of an individual's true score (see the discussion of *standard error of measurement* in Chapter III). The error in prediction for a particular individual is unknown. With the use of the generalized standard error of estimate, however, the counselor can establish the gross probable limits of his misses in prediction for particular counselees. This range may be fairly wide and will often limit the predictive usefulness of the measure.¹⁰

Except when extremely high or low scores are obtained, the counselor can find little in test manuals that can be used with precision in answering questions raised by counselees. If, in using a test in which the reported validity is $r = .60$, the counselor finds that one of his counselees scored at the 50th percentile, he will realize that the expected range for this person on the criterion measure lies somewhere between the 21st and 79th percentiles two times out of three!¹¹ This rather unspecific, unprecise prediction should make the counselor humble and wary of being dogmatic in pronouncements of probable future success or failure of a counselee that are based mainly on a test score.

To compensate for this wide band of error, some increase in the practical usefulness of test scores may be obtained if the counselor is concerned with a very simple and crude criterion measure. In general the coarser the criterion measure, the greater assurance there will be that the prediction may be relatively accurate. Conversely, when more exact and precise predictions are attempted, the expected error in prediction will be greater.¹²

In some practical situations the counselor may be satisfied with a relatively coarse estimate of success. He may wish to use a two-

¹⁰ Guilford, *op. cit.*, pp. 377-384. Also George E. McCabe, "How Substantial Is a Substantial Validity Coefficient?" *Personnel and Guidance Journal*, February, 1956, 34:340-344.

¹¹ This is not, however, a statement of probability in the technical sense. The statement that his true score lies within the limits is either true or false and not precisely subject to the two out of three statement. The use of gross limits as they are used here seems justified in general use even if it is not technically quite accurate.

¹² Clark Hull. *Aptitude Testing*. Yonkers, N.Y.: World Book Co., 1928, p. 267.

point scale such as pass or fail, or he may decide to use a simple three-point scale. If such coarse criteria will serve his purposes, he will find that interpretations can be made with a *greater degree of assurance* than if he were working with such matters as grade point averages, rank in class at graduation, units of production, future sales in dollars or precise percentile standing. When test scores are to be used for selection purposes and a fairly coarse criterion is used, the percentage improvement over chance of even a moderate coefficient becomes more useful. The following example from a bulletin of The Psychological Corporation illustrates the case.

What permits us to use tests effectively even though their validity coefficients are considerably lower than .866? First, there is the *matter of precision*. The standard error of estimate refers to the band of error around predictions of precise, specific rankings of each individual on the criterion. *In most practical work, such precision is unnecessary*. We do not ordinarily need to predict that John Jones will be exactly at the 85th percentile in a college class, or that Bill Smith will be 19th in a group of 25 engineering apprentices. We are far more likely to be concerned with whether Jones will survive the first year in college, or whether Smith will be one of the satisfactory apprentices. For these purposes, whether Jones is at the 75th percentile or 90th percentile is of lesser moment; we can make a quite confident prediction that he will succeed, even though there may be a fair-sized standard error of estimate applicable to the specific percentile our formula predicts.

A second factor working in our favor in the practical use of tests is that, as the opening quotation notes, *predictions are most accurately made at the extremes*—and it is the extremes that are of greatest interest to us. Few colleges grant large scholarships to more than 10 or 20 per cent of their students. Few colleges fail as many as half their students and few industrial firms fire as many as half of those they hire. More often, the failures are 10 per cent or 20 per cent or possibly 30 per cent—the extremes. Thus a test which does not predict with accuracy whether students will be at the 40th percentile or the 60th percentile, can still do a valuable service in predicting that

very few of the high scorers will be in the 20 per cent who fail during the freshman year, or that hardly any scholarship winners will be academic failures. In industrial selection, a test of moderate validity can be efficient in quickly screening out the "clearly ineligible" from the "clearly eligible." There will remain an indifferent zone of test scores for persons in the "eligible" range; for them, other considerations than test scores may determine whether they should be hired.

Let us look at some data. One hundred ninety-one eighth-grade boys took the Verbal Reasoning Test of the Differential Aptitude Tests (DAT) battery at the start of a term. At the end of the term, the grades they earned in a Social Studies course were obtained. Seventy-six were found to have earned grades of D or lower; they represented 40 per cent of the total class. On the basis of chance (i.e., using a test with zero validity), we should expect to find that 40 per cent of those at each test score level—low, medium or high—obtained grades of D or lower. The coefficient of correlation between the test scores and these grades was .61, for which the index of forecasting efficiency comes out to just 20 per cent better than chance—hardly enough to notice. Table [21] reveals a very different story—it shows the test to

TABLE [21]. Chance Expectations and Actual Performances in a Social Studies Class in Relation to DAT-Verbal Reasoning Scores

DAT Verbal Reasoning Test Score	No. of Pupils	% Expected by Chance to Earn D, E, or F	% Actually Earning D, E, or F
26-up	19	40	6
18-25	49	40	14
10-17	60	40	36
2-9	63	40	73

be a highly efficient predictor for the school's purposes! Instead of 40 per cent of the highest-scoring pupils being found in the low grades group (as one would expect by chance), only six per cent are found there.¹³ (*Italics added.*)

Despite the remarkable improvement over chance illustrated

¹³ Alexander G. Wesman, "Better Than Chance," *Test Service Bulletin*, New York: The Psychological Corporation, May, 1953, pp. 9-10.

above with a three-point criterion scale (D, E, or F), the counselor has the task of estimating whether or not his *particular* counselee will be among the high-scoring 6 percent who will still receive poor grades, or among the bottom-scoring 27 percent who did not. He is faced with the problem of attempting individual prediction from statistics based upon the average performances of groups. This is a real dilemma. He must preface his answer to his counselee's question about his probable success in college with ". . . that depends."

SELECTION OR COUNSELING? When test scores are used for *selection* purposes rather than *counseling*, the problem may be one of determining which of many applicants for training are most likely to succeed. Another selection situation might require the estimation of the grade point averages for all members of an entering class of college freshmen. In cases such as these the admissions officer or personnel manager is concerned with the accuracy of prediction at *any* degree above chance. If his predictive equation misses a few cases, the failure may be dismissed as chance error, and he may be satisfied that he is more often right than not. And in certain instances the percentage of misses, as in the illustrations above, may be small. Even if these misses are but one, ten, or 20 out of 1,000, they are of much greater consequence to the counselor than to the admissions officer or personnel manager. For the counselor has deep personal concern for *each* of the 1,000 persons. His major responsibility is to the optimum development of each individual counselee.

The counselor who has followed the discussion above must realize that, since most test manuals report validity coefficients in the range of .40 to .60, predictions based upon scores from tests improve pure guessing by approximately 10 to 20 percent *on the average*. How can a counselor improve his predictions? What information needs to be combined with test scores in order to increase predictive efficiency? What additional evidence of performance of the counselee needs to be weighed in attempts to answer

the questions raised by counselees, their parents, and their teachers? What must be considered when the counselor says ". . . that depends . . .?"

INFORMATION PROVIDED BY PERSONAL DATA

Every experienced counselor has worked with counselees for whom test scores do not seem to tell the whole story. The student with a marked reading difficulty, for example, may be handicapped on many standardized tests. With elementary school pupils it becomes difficult to determine whether their intelligence test scores are low because they cannot read well or cannot read well because they lack "intelligence." Study of clues from the pupil's home background, his everyday performance in class, his social class status, or his overall behavior may assist in solution of the problem. Some students test high but achieve relatively poor success in their academic attempts. Others do the opposite. Prognosis for success of some students in post high school training or in certain occupations may be determined not only by their test scores and school achievement but also by the pressures of their families. For some pupils such pressure produces a high state of tension that reduces their efficiency to a marked degree.

Jerry is a case in point. His father sought counseling for him after he indicated that he was uninterested in the West Point appointment his father had secured for him. His father was insistent that Jerry enter training for engineering because that was a "most respectable and well-paying profession." Jerry scored high on the usual preengineering aptitude tests, which included subtests dealing with mathematics, mechanical comprehension, and spatial relations. On the basis of such scores the counselor's approval of a choice of engineering training seemed warranted. Because of conflict with his family over the father's choice of engineering, Jerry refused to consider it. He did express a desire to go into the field of journalism or professional

writing. His leisure-time activities and stated interests were more in line with a vocational goal in literary areas than in engineering. By full exploration of various background factors with Jerry, his teachers, and his father the counselor could make a more meaningful interpretation of his test scores in relation to his vocational choice.

TEACHERS' EVALUATIONS

Teachers' marks are still the unreliable, invalid, but indispensable evidence of success in our schools. They are, and will probably continue to be for many years, the "coin of the realm" in our schools. As such they are important supplements to test data in the appraisal of current and future achievement of students. High school grades are strongly influenced by teachers who tend to reward children of the higher social classes, give better marks to girls than boys, and become influenced by such factors as effort, neatness, or "apple polishing."¹⁴ It should not come as a surprise, despite the unreliable nature of grades, that test scores are sometimes not as efficient predictors of grades in college as previously earned high school grades.

Grades contribute a dimension that "on the spot" testing fails to provide. They may provide a longitudinal frame of reference with which to view the counselee's total performances. The four-year record of marks of a high school senior can furnish important information for the counselor to supplement the testing record. As teachers' reports become more analytical and diagnostic, their value in portraying the typical daily behavior of students will be greater than that of the summation type of grade.¹⁵

¹⁴ See: Harl Douglas and N. Olson, "Relations of High School Marks to Sex in Four Minnesota High Schools," *School Review*, April, 1937, 45:262-288. Robert S. Carter, "How Valid are Marks Assigned by Teachers," *Journal of Educational Psychology*, April, 1952, 43:218-228. William S. Learned, *What's in a Mark?* The Carnegie Foundation for the Advancement of Teaching, Thirty-Seventh Annual Report, New York: The Carnegie Foundation for the Advancement of Teaching, 1940, 36 pp.

¹⁵ Eugene R. Smith and Others, "Appraising and Recording Student Progress,"

Testing provides a sample of behavior within a highly structured and somewhat artificial situation. Longitudinal analysis of the counselee's normal operation in a functional situation made by observing teachers may bring to light characteristics completely missed in testing. These may include work habits, social skills, attitude toward schooling, values, and level of aspiration as well as a host of others. As more teachers become well trained in observation, child study, and diagnostic procedures, psychological testing may become relegated to even a lesser role than it has now in the process of counseling. The experienced counselor is more apt to give more weight to a capable teacher's estimate of a student's performances than to his test scores despite the former's lack of quantification.

Counselors are aware that grades are often rewards for compliant behavior, accuracy, neatness, dependability, and willingness of the student to assume the teachers' outlook and values. Viewed in this manner, grades may provide estimates as to how the counselee will function in some future situation in areas not tapped by testing. Clues and hints as to his reaction to authority figures, his ability to "adapt" to the realities of a situation, his typical canalization of hostility impulses, or his social skills in playing the "rules of the game" may be gained by study of grades and teachers' observations in relation to test scores.

In addition to recognition of the fluctuation of student performances reflected in his marks, the counselor should also be aware of teachers' constant errors in assigning marks. Suggestions as to how this constant error factor might be quantified and appraised are given elsewhere.¹⁸ This procedure, which normalizes each teacher's grade distributions, assigns a weighting factor to the grades they award, and allows the counselor to evaluate the given teacher's

Adventure in American Education, Vol. 3. New York: Harper & Bros., 1942. John W. M. Rothney, "Evaluating and Reporting Pupil Progress," *What Research Says to the Teacher*, No. 7. Washington, D.C.: National Education Association, 1955.

¹⁸ John W. M. Rothney and Bert A. Roens. *Counseling the Individual Student*. New York: Dryden Press, 1949, p. 205.

marks in relationship to all marks given in the school. In this manner he has a useful way of understanding the fact that an "A" from Miss Brown is given as often as a "B" in Miss Smith's classes and consequently has less suggestion of excellence than an "A" grade from Miss Smith. In knowing some of the possible variables that lie behind the awarding of marks by teachers in his school, the counselor is in a better position to appraise the meaning of his counselee's grades in relation to future academic requirements and courses.

The following examples from counseling files may further illustrate the interaction of grades and test scores.

SHELIA

Shelia, a freshman in college, appealed for counseling to see "what was the matter" with her current classroom performances since she was receiving D's and F's in nearly all her college courses. This record came as a distinct shock to her because she had graduated fifth in a class of 156 high school students, and had always thought of herself as an above-average student.

Her test record follows:

<i>Test</i>	<i>Year</i>	<i>Score</i>
Otis, Gamma	10th grade	IQ 110
Ohio State Psychological	12th grade	65 percentile
Kuhlman-Anderson	12th grade	IQ 117
Coöperative English	13th grade total	90th percentile
American Council on Education Psychological Examination for College Freshmen	13th grade	Total 50th percentile

Examination of her test record alone seems to indicate that she would succeed in college level work. The ACE score is somewhat lower than the other scores, but this could be laid to the

stress of freshman week and the tensions and excitement of the period during which the test was administered.

Following are examples of what her high school teachers had to say about her classroom performances:

English, 12th grade: "Shelia is a good student who always hands her work in *on time*, and seems to do more than is expected of her. She is a pleasure to have in class for she seems to love school and always *tries her best*. I can always count on her to come through no matter what the assignment."

English, 11th grade: "Shelia *tried very hard to please*. I sometimes think she has the reputation for being a good student simply because she gives such a *good impression*."

Social Studies, 12th grade: "This girl does much better work when she is following a definite assignment. She seems lost when she has to dig out reference materials in the library and *needs a great deal of help when asked to work on her own*."

High School Counselor: "Shelia has had a lot of pressure from home to achieve. Her mother is constantly worrying about her grades. She is very determined not to let her mother down and *works long hours* to get her work done. She is going on to college *without any definite vocational goals* and seems unable to make a vocational decision lest it displease her mother."

College Counselor: "Shelia is greatly troubled about the pressures from home. She stated in her initial interview with the college counselor that she was afraid and lost at college. She seemed unable to proceed alone and unguided as she had not been required to do in the more protective high school setting. She was somewhat at a loss on what procedures were available to make a personal impact on her instructors. Her promptness and compliancy, which paid off in the high school setting, did not seem to be effective in the college situation, and she was near panic because of this. She was studying until midnight nearly every night and week-end, but found little reward from her continual efforts to compensate for the lack of quality of work with sheer quantity.

The counselor tried to help her to see that, while she had an

average amount of talent as measured by various tests of academic promise, she would be unable to succeed in college with the same techniques of pleasing teachers, doing extra work, or just working harder and harder as she had done in the high school. She was given help in developing new study habits and shifting her level of aspiration somewhat downward in an effort to provide her a more realistic understanding of herself.

Why the high school counselor did not temper his evaluations of Shelia by careful appraisal of her high school teachers' qualitative assessments (each of these teachers had given her A's) is not known. Certainly this girl should have had a more realistic degree of self-appraisal before entering college where she was forced to meet reality. Her overall grades were very high, but a closer analysis of her actual work habits and techniques for getting good grades while in high school would have supplemented knowledge of her potential performance in conjunction with her high test record. Analysis of the disparity between her *ACE* score and her *Coöperative English* score pointed to specific strength in English grammar, but also provided evidence of her average performance when the task was to generalize and abstract this learning.

Another example of a counselee whose test record and academic achievement need qualitative analysis because of the various factors involved in teachers' evaluations is given below.

JEFF

Jeff requested counseling from a college Guidance Center during the summer following high school graduation, and prior to his enrollment in college. He was concerned about his inability to earn above average grades while in high school, but upset because his high school teachers and counselors had always encouraged him to go to college despite this mediocre academic record. They had told him that he had "lots of ability, if he would only use it."

His rank in class at the time of graduation was 378 out of a

class of 459 students. His mother called the college counselor and expressed her anxiety over his chances for success in college because of his low scholarship. She expressed her own hopes that he might succeed. She stated that she had been told by the high school counselor that he did much better on tests than he did in classwork.

His testing record follows:

<i>Test</i>	<i>Year</i>	<i>Score</i>
Otis, Beta	7th grade	IQ 115
Otis, Gamma	9th grade	IQ 125
California Achievement Test (Reading)	11th grade	13.6
Differential Aptitude Tests	11th grade	
Verbal Reasoning		90th percentile
Abstract Reasoning		85th "
Numerical Reasoning		50th "
Space Relations		15th "
Mechanical Comprehension		25th "
Language Usage		
Spelling		35th "
Sentences		30th "
Wechsler Adult Intelligence Scale	between 12th and 13th grades	Verbal IQ=124 Performance IQ= 118 Total IQ=123
American Council on Education Psychological Examination for College Freshmen	13th grade	Total 85th per- centile

The high school counselor's summaries of teachers' evaluations and his own analysis of Jeff's academic performances follow:

Freshman year counselor's report: "Jeff's appraisal of himself is very realistic. He states that he *lacks confidence* in himself and

is *very shy*. He is not popular with the girls but enjoys a certain popularity with some of the boys because he is a good fellow and is willing to undergo their ridicule and laughter. His appearance is against him. He is tall and very thin and wears very thick glasses."

Sophomore year counselor's report: "Jeff has been in to see me a number of times this year at the request of nearly all of his teachers. They report he 'cuts up' in their classes and *does not get his work in on time*. His library privileges have been withdrawn because of his loud talking and giggling in the library despite repeated warnings. He is the 'clown' of his class and is constantly seeking attention in classes by *talking out of turn*, making unnecessary and sometimes *rude remarks to his teachers*, and generally 'getting in trouble.' "

Junior year counselor's report: "Jeff has never learned to work well in classwork. He finds it easy to stay home and *misses school frequently*. His parents are planning a college career for him, but he seems only *mildly interested*. The father is very much discouraged because he knows that Jeff's academic record will preclude his entrance to any high status college."

Senior year counselor's report: "Jeff seems to take pride in boasting that he '*did nothing and got away with it . . .*' I believe that this is a defense mechanism and he is not fooling even himself. In class he is *very quiet*, but on occasion is *surly and rude* when requested to do his classwork by his teachers. He seems now to have a '*chip on his shoulder*' all the time."

(In both of the illustrative cases key phrases have been underlined in writing the reports to accentuate the attitudes, skills, and behaviors that were crucial in evaluating the grades these high school students received.)

Evaluation of Jeff's academic performances made at the time of his high school graduation would have indicated that he was not likely to succeed in college level studies. When Jeff sought counseling from a college Guidance Center midway through his first year, he was failing in nearly all his college course work. By the middle

of his second college year he was doing average work, largely because of the efforts of his counselor, who helped him to change his attitude toward his father and other authority figures. Many hours were spent with Jeff in the Reading Clinic to remedy his poor study habits and to prepare him better in terms of needed skills to attack his college course assignments. He lost most of his belligerent attitude as he clarified his vocational goals and stopped fighting the world. Relieved of some of the pressures and tensions that had troubled him all through high school, he was able to use his talents to better advantage and achieve some measure of success. Looking back, his college counselor was able to evaluate Jeff's poor grades in high school in terms of poor study skills plus a negative attitude toward school work.

These case examples clearly point out the need for a careful *qualitative* analysis of the grades a student receives in high school. Shelia received good grades in part because of her compliancy and willingness to agree to her teachers' demands. Jeff was penalized without regard for his level of talent by his refusal to do so. In each case quantitative analysis of the grades or the test performances alone would not have told the whole story. Only by careful interpretation of the test scores plus understanding of the factors behind the grades received was the counselor able to aid these college students in development of a realistic picture of themselves.

SOCIAL CLASS STATUS

Considerable attention has been paid in the past few years to the testee's social class status as a factor that must be considered in the interpretation of his test behavior.¹⁷ For a long time educational psychologists acknowledged that more "brain power" was to

¹⁷ See: Kenneth Eells and Others. *Intelligence and Cultural Differences: A Study of Cultural Learning and Problem Solving*. Chicago: University of Chicago Press, 1951. William N. Leonard, "Psychological Tests and the Educational System." *School and Society*, April 12, 1952, 75.225-259. Walter B. Brookover, "The Implications of Social Class Analysis for a Theory of Education," *Educational Theory*, August, 1951, 1:97-105.

be found among children of higher socioeconomic positions on the theory that those families which had accumulated property and position gained it because of their greater native intelligence. It was said that they kept their positions because their children also had superior brains and achieved superior test and scholastic performances. The future success in the world of children from these families was assured, so this argument went, but the acquisition of family wealth, position, and power was reinforced by their superior intellectual endowment. Recent analysis of test scores of members of varying social classes, as undertaken by Eells¹⁷ and others, suggests that the differences between scores of children of different social classes may be more a function of the test items than inherited intelligence. While this point of view has not been fully accepted, it has created enough sensitivity to this issue to make the consideration of a counselee's social class status more important than before in the evaluation of his performance on tests.¹⁸ A typical study comparing the performance differences of social class position with respect to school grades and performance on the Henmon-Nelson Tests of Mental Ability showed group differences above the 1 percent level of significance and in favor of middle classes over upper-lower classes.¹⁹

The counselor who works with a student from other than a middle-class home might question the validity of a score derived from a test that was based on middle-class constructs. He might conclude that a superior test performance achieved by a counselee from a lower social class was even more meaningful than it seemed when it was compared with general norms. In some situations a counselor might find it necessary to develop social class norms as aids in his test interpretations. He will certainly want to gather more data about the individual counselee than those provided by scores on group paper-and-pencil tests that are highly weighted

¹⁷ Quinn McNemar, "Review of Intelligence and Cultural Differences." *Psychological Bulletin*, July, 1952, 49:370-371.

¹⁹ Robert A. Heimann and Quentin Schenk, "Relations of Social Class and Sex Differences to High School Achievement." *School Review*, July, 1952, 49:213-221.

with verbal items. Such items may penalize a student from a lower-class home.

Still another variable, the cost of higher education, needs to be integrated into the total picture before the statement, ". . . that depends . . .," can be completed. It has been shown that more than 40 percent of the top quarter of superior high school graduates do not go on to post high school training. There is evidence to indicate that, with the exception of the very superior students, there is a high relationship between college attendance and the socioeconomic level of the family unit.²⁰ Some studies²¹ indicate that nearly four times as many college youth come from homes of professional parents as from homes of farmers and laborers.

It seems clear from the above that the social class variable must be considered when the high school counselor is trying to help a student to estimate his chances of success in post high school training.

The value systems of middle and upper classes seem more in tune with the demands of academic schooling. This is particularly true of the concept of vertical social mobility through education.²² In middle and upper social classes the motivation for success within the framework of the school system has certain aspects of identification with the teacher. Students of higher social classes are frequently members of the same social group that creates and maintains the values taught in the classroom. Upper-class students may thus have an advantage in the accumulation of higher marks,

²⁰ Byron Hollingshead, *"Who Should Go to College?"* New York: Columbia University Press, 1952.

²¹ See: Ralph Berdie, *After High School, What?* Minneapolis: University of Minnesota Press, 1954. L. T. Phearman, "Comparison of High School Graduates Who Go to College with Those Who Do Not," *Journal of Educational Psychology*, November, 1949, 40:405-414. Dael L. Wolfe, *America's Resources of Specialized Talent*, Report of the Commission on Human Resources and Advance Training. New York: Harper & Bros., 1954.

²² See: August Hollingshead, *Elmtown's Youth*, New York: John Wiley & Sons, 1949. W. Lloyd Warner and L. Srole, *The Social Systems of American Ethnic Groups*, Yankee City Series, III. New Haven: Yale University Press, 1945. Theodore Caplow, *The Sociology of Work*, Minneapolis: University of Minnesota Press, 1954.

which influence the awarding of scholarships and meeting college entrance requirements. Their records may reflect the fact that teachers tend to identify with, and accept more readily, those students whose backgrounds encourage learning of the middle-class curricula taught in many schools.²³

Recognition of possible bias in test performance owing to cultural handicaps should make a counselor careful in the interpretation of scores in relation to vocational or training goals. A counselee who comes from a distinct subculture or a minority group needs more than casual consideration, owing to the strong possibility that his test performance may present a distorted picture of his real potentialities. Adequate tests and adequate interpretative norms are lacking for many of our minority groups, who may be unduly handicapped in seeking vocational goals by their low test performances.

The counselor must recognize the possibility of cultural bias in operation as he works with students from all classes on problems of vocational choice. In some cases this may mean helping a student who comes from a lower-class family, but who has opportunity to adjust his vocational sights upward. In other cases it may mean aiding a high school student from an upper-class home to consider a more realistic vocational choice than that sought by his family. In schools with a large group of Spanish-American, Indian, or other bicultural groups and first generation children of foreign-born parents, counselors can anticipate that many of the students will get low test scores. This is particularly true in the early grades. Many of these lower-class students drop out of school before high school graduation. In many cases they do not even enter high school, or if they are forced to do so by compulsory attendance laws, they leave as soon as it is legally possible.²⁴

²³ W. Lloyd Warner, Robert L. Havighurst, and Martin B. Loeb. *Who Shall Be Educated?* New York: Harper & Bros., 1944.

²⁴ National Manpower Council. *A Policy for Skilled Manpower*, New York: Columbia University Press, 1954.

The answer to this loss lies in more than just adequate counselor appraisal. It demands complete restudy of the school's aims and curriculum by the school and community. Until this happens the counselor is in the best position in the school to help identify potentiality in the early school grades. He can encourage youth of lower social classes to stay in school until graduation and help them to plan their post high school training. Teachers and counselors who are interested in more than the development of compliant behavior and mass conformity can help to identify talented children of *all* social classes very early in their school careers. Encouragement and assistance over a period of several years are necessary to aid them to raise their vocational aspirations above the usual pattern of early school leaving and quick entrance to the labor market.²⁵

MOTIVATION AND OTHER INTRAPERSONAL FACTORS

The study by Hollingshead mentioned above revealed that many youth do not consider college training seriously because of unfamiliarity with and lack of acceptance of the contribution of education to vocational and personal satisfaction. Counselors who work with high school youth are familiar with the talented youth who has gained little satisfaction from the curriculum of his school, and who persists in school only because it provides a diploma. His attitude toward schooling is simply that it must be borne and tolerated until graduation releases him into the "real" world where jobs, pay envelopes, and cars provide real satisfaction.²⁶

Some demands of teachers seem to inculcate in their students work habits that become handicaps in the taking of tests. The student who has learned, at the insistence of his teachers, to work slowly and methodically and to check carefully for mistakes is

²⁵ Harry Belin, "The Utilization of High Level Talent in Lower Socio-Economic Groups," *Personnel and Guidance Journal*, November, 1956, 35:175-178.

²⁶ See: the case of Clark in John W. M. Rothney. *The High School Student*, New York: Dryden Press, 1954, pp. 73-80.

likely to get so few test items completed that his score will be low. While *this behavior may be significantly important in the pursuit of certain careers or of some classwork*, it tends to be a handicap in test-taking. This is especially true when subtests have time limits of three to six minutes and are designed to measure speed rather than power. In such cases the teachers' analytical judgments of the student's usual behavior may be important in interpreting the test score.

Test authors seldom pay much attention to the problems of motivation in their manuals although it is known that performances can be influenced by the attitudes of the subject toward the testing situation. In a school visited by one of the authors, senior students confided that they had "faked" low test scores on their placement tests so that they would be placed in lower or easy sections. They knew that they could make better marks with less effort in such classes. Some research²⁷ suggests that when the counselee participates in the choice of tests, motivation in the testing situation is greatly improved. In mass testing programs motivation is frequently low because the counselee fails to realize that the scores may give him some important answers in terms of his goals and decisions. Too many schools and too many teachers have rejected testing, too, because they have seen situations in which the tests were given and the results securely filed away out of sight of everyone concerned. If, in the interests of economy of time and effort, *an administrative decision is made to test a large group of students at one time (and such action is more the rule than the exception)*, attempts should be made to get all students motivated. Group guidance classes, home room discussions, assembly talks, and films may be utilized to awaken interest on the part of the testees by helping them to see how the forthcoming test results might be useful in their plans. If clear recognition is given to the procedure by which each student will be apprised of his test per-

²⁷ Ray H. Bixler and Virginia H. Bixler. "Test Interpretation in Vocational Counseling." *Educational and Psychological Measurement*, Summer, 1946, 6:145-55.

which there is uncritical acceptance of test results in a fixed, dogmatic, and rigid manner. With such counselees test results, improperly assessed or inadequately interpreted by the counselor, may cause sharp alterations in their plans and even personal disorganization as they attempt to translate their scores into behavioral constructs. The counselor has an important responsibility in the interpretation of test scores for these persons because they do not have enough technical competence or toughness of mind to regard test performance with proper skepticism. This unqualified and uncritical "digestion" of test results may do a serious disservice to such persons. Counselors need to assess this tendency to "test trauma" before they attempt to interpret test scores to counselees.

In the making of decisions involving vocational and educational choices, the inter- and intrapersonal factors of behavior often seem to be as important as the intellectual and academic. At times the decision needs to be resolved in terms not only of what a particular counselee can achieve but of how he functions in relation to other persons. Industrial psychologists frequently report that unsatisfactory performance on the job is more a function of personal-social adjustment than intellectual skills²⁹ and it has been shown that teachers' grades may be influenced by interpersonal relationships. The counselor will, therefore, need to assess a student's behavior in the personal-social area as well as in intellectual achievement.

Counselors usually agree that data concerning the typical behavior of counselees' interaction with other persons are essential in guidance. General recognition is given to the fact that emotions, feelings, attitudes, and personal values are of great significance in vocational counseling and counseling psychologists have given increasing emphasis to the development of realistic self-images by counselees. They would like to incorporate added insight into these

²⁹ See: R. Prator, "The Employer Survey and General Education." *California Journal of Secondary Education*, November, 1950, 25:438-440. H. C. Hunt. "Why People Lose Their Jobs or Aren't Promoted." *Personnel Journal*, 1936, 14:227.

areas as well as the more common "aptitudes and abilities" in the total vocational counseling process. They are not in agreement, however, about how such insights may be obtained. Many turn to highly structured instruments, others use unstructured methods, and some depend largely on descriptions of behavior by observers.

Evaluation of structured and unstructured devices is discussed in detail in Chapter VIII. The following section will deal with one of the descriptive techniques that may be used effectively in gathering information about the behavior of counselees.

BEHAVIOR DESCRIPTION

One promising technique in the assessment of functioning in the area of personal-social behavior is the Behavior Description Method developed by the Records and Reports Committee of the Progressive Education Association in its Eight-Year Study.³⁰ This instrument was designed in an effort to get away from the more common "personality" tests, rating scales, and check lists of behavioral traits. It attempted to provide a ". . . definite procedure for studying attitudes, habits and traits, with a technique for reporting them and recording the results."

It has been amply demonstrated that this technique, called *The Method of Behavior Description*, can provide valuable evidence of pupil's development in the areas for which coverage is often attempted by rating scales, personality inventories, checklists, and anecdotal records. The entries in this "behavior description method" resemble those in checklists and rating scales, but there is a basic difference—the users' belief in the importance of individuality and of individual differences. Those who plan to use the method must prepare carefully defined descriptions of the pupil's behavior, which may, sometimes, resemble those used in rating scales. Class-

³⁰ See: Progressive Education Association, The Reports and Records Committee. *Manual for Behavior Description*, 1937. John W. M. Rothney and Bert A. Roens. *Counseling the Individual Student*. New York: Dryden Press, 1949, pp. 96-105.

room teachers, and others who have had sufficient opportunity to observe the pupil, place symbols indicating their relationship to the student beside the description that best fits him. The accompanying chart illustrates the method.

Pupil: Mary Anderson—Junior High School			
Descriptions	Grades		
	7	8	9
Appears to feel secure in and is accepted by groups of peers	E, MU		
Appears to feel anxious about her standing in her groups		MU, E	MU, E
Wants to belong to groups but is generally treated with indifference		HE	HE
Withdraws from peers so much that she is not fully accepted			
Characteristics of her person or behavior cause rejection by her group			
(E—English; MU—Music; HE—Home Economics)			

This description indicates that Mary felt secure and was well accepted in English and music groups in Grade 7, but had begun to show some anxiety about relationship to her peers in those classes in Grades 8 and 9. Something happened to cause the girls in the home economics class to treat her with indifference in two upper grades. A classroom teacher who wanted to help Mary would seek to uncover the events leading to the difficulty. In addition to the abbreviations placed opposite the descriptive items, some teachers will want to add supplementary notes or explain what lies behind the descriptions they have given.

The differences between this descriptive procedure and rating is

just that the describers try to *summarize* what has been observed while raters attempt to *judge* the quality of the observed behavior. There is no implication in the "behavior description method" that any particular kind of behavior is best for one child at a particular time. The technique admits the well-known fact that the child's behavior may vary in different situations and under changing influences. Thus, though each reporter makes a correct description of what he observes, the reports about an individual may differ greatly at any given time. The procedure allows for the possibility that differences in the descriptions of various observers may be as significant as the similarities they report. It must be emphasized that there is no implication of goodness or badness in the use of the term "behavior."

Instead of requiring a perfunctory rating of personality twice a year, a practice that classroom teachers dislike and if possible avoid, the behavior description method proposes that teachers be encouraged to make continuous observations of their pupils with respect to the defined characteristics and to record their descriptions at such times as are decided upon. Duplicated sheets of the definitions of characteristics are furnished to the teachers so that they can make their descriptions of the pupils with the definitions before them, and without being influenced by each other's observations. The descriptions can be entered on sheets of class lists with the characteristics used as headings across the top of the page. Abbreviations and numbers for types make such a form simple to prepare. The descriptions are transferred from the class list to the central record card, thus making a picture of the pupil as seen by all his teachers. When significant notes accompany a description, they can be entered on the record card beside the definitions. If teachers study the form and the definitions of behavior at the beginning of the school year and agree to make the descriptions upon the basis of carefully considered evidence, the descriptions are likely to be valid.³¹

³¹ John W. M. Rothney, "Evaluating and Reporting Pupil Progress," No. 7, *What*

With the information provided by the behavior descriptions the counselor can take another look at his test scores and reevaluate them. When they are further supplemented by teachers' grades, social class data, and information on interpersonal factors, test scores can be interpreted more meaningfully.

CONTRADICTORY EVIDENCE: CASES

In this chapter it has been suggested that the process of using test data in counseling is a complex procedure. Many ideas run through a counselor's mind as he looks over the test and academic performances of a given counsellee prior to the counseling interview. The interplay of test data with other forces in the process of helping a student reach decisions is illustrated in the following cases.

ED

Ed was a high school senior who wanted to know if he should go to college after he was graduated from high school. He ranked 33d in a graduating class of 100 students. In six semesters he received 6 A's, 23 B's, 6 C's, and 1 D. His test performances are recorded below.

The counselor might assume on the basis of the test record that there was a poor chance for success in college for this boy whose average percentile rank was low. Despite this relatively low level of performance on psychological tests, Ed's scholastic record indicated that he was consistently regarded by his teachers as an able student. (Which predictor does the counselor use in counseling with Ed? Should he use test scores that indicate little chance of success in college or data obtained from cumulative teacher's marks which seem to indicate average or better chances for success? To

10th Grade	Percentile	11th Grade	Percentile
Henmon-Nelson, Test of Mental Ability	10	Henmon-Nelson, Test of Mental Ability	16
SRA Primary Mental Abilities		Differential Aptitude Tests	
Verbal	15	Verbal Reasoning	45
Space	20	Number Ability	10
Reasoning	55	Language Usage	
Number	30	Spelling	0
Word Fluency	10	Sentences	25

what extent does the counselor use both sources of data and combine them?)

Five years after graduation from high school Ed was completing his senior year in a Big Ten University where he majored in art. Although he was making adequate grades in his final year, he had some trouble during his first year. At that time he dropped out of school and went for one year to a small teachers' college. (What other factors in Ed's cumulative record should have been added to the data at hand in order to make meaningful interpretation? Is Ed an example of the "overachiever," or is he the type of counselee for whom tests have little personal validity?)

Ed's family had exerted a great deal of pressure on him to go to college and become a professional person as had all others in his immediate family group. During this period of pressure they discouraged his tentative vocational choices, which included ceramics, woodwork, owning and operating a hobby shop, or "just working with my hands." Subsequent statements of vocational choice found Ed more and more vague as to just what he would do after high school graduation, but he did say that he probably would accede to his family's wishes and go to college.

Some anxiety was shown in his hesitation about arriving at even a tentative vocational decision and in his increased dependence on the counselor. Discussion of occupational information became difficult because Ed became too eager to obtain and follow suggestions from the counselor.

Ed's teachers reported that he tried very hard to please them, and one described him as the "Boy Scout type," eager, pleasant, helpful, and compliant to a great degree. They reported, however, that he was a poor reader. This was an important weakness, which the counselor and Ed considered carefully in interpretation of his test scores and in estimating his chances for success in post high school training. Several teachers had reported that Ed was a slow, methodical worker who was overconscientious in his desire to be right. He would not guess at any items. This set in test-taking seriously inhibited his performance on the tests that had time limits and made heavy demands on rapid reading.

In summary, the counselor could not possibly give Ed a brief yes or no answer to his questions about possible success in college on the basis of his test scores alone. Into the equation dominated by low test scores it was necessary to add the better-than-average school grades, the reading handicap, and his anxieties and indecisions. Perhaps the whole equation would need further modification in the light of Ed's lack of clear-cut reasons for going on to college.

Perhaps the one problem referred to the typical high school counselor with the greatest frequency by teachers is the case of the pupil who is reported as "not working up to capacity." (Obviously not many persons work to capacity regularly or mental hospitals would be even more crowded than they are at present. Continuous working to the limits of capacity would probably harm the mental health of even the staunchest subject matter specialist.) Instructors are confronted almost daily with the student who tests high, but does less than average work in his classes. This fact causes many teachers to become ego-involved. They can't understand why the student does not develop the same deep love as they

have for the "Lady of the Lake," the War of 1812 and all its battles, or the rules for the use of the comma. This is often true of the students who have achieved high test performances. It seems to the teacher that such students are willfully doing poor work, and they may frequently exhort him to "work up to capacity." Bert was an example of this type of high school student.

BERT

Bert came from a lower-class family, which seemed proud that they had all quit school before graduation and secured well-paying jobs in factories and on railroads. Bert was breaking with family tradition by persevering in high school and announcing his intentions to stay until graduation. During the three years of counseling with him, Bert's counselors had suggested that he might profit from college attendance. When he decided to go to college the family scorned him. They were disturbed that one of their members planned to go to school for four more years rather than getting a job and bringing home a pay check.

Bert, who was the smallest senior boy in his class, had worked out his particular mode of adjustment to the world early in his school career (before making his decision to go to college) by becoming the "worst behaved" boy in school. His teachers did not even turn around from writing on the blackboard when they heard a disturbance in class. They simply remarked, ". . . Bert, be quiet!" His reputation as a "cutup" and "wise guy" was firmly established and Bert proceeded to "get by." He did not obey the rules of the school or get his assignments in on time. He indicated that he was uninterested in the class and that he saw no need to cooperate with his teachers. As a result Bert's academic record showed 3 B's, 18 C's, and 8 D's over the last three years of high school, and he graduated in the lower third of his class.

Bert's test record belied this academic record. Of the 22 tests that he took in his last three years of school, as shown in the following chart, only two dropped below the 85th percentile.

Tests	Percentile 10th Grade	Percentile 11th Grade	Percentile 12th Grade
Henmon-Nelson Tests of Mental Ability	95		
SRA Primary Mental Abilities			
Verbal	90		
Space	66		
Reasoning	90		
Number	91		
Word Fluency	91		
Coöperative Reading			
Vocabulary	89	65	84
Speed	89	96	97
Comprehension	89	87	92
Total	89	87	92
Differential Aptitude Tests			
Verbal Reasoning		80	
Number Ability		95	
Spelling		95	
Sentences		95	

When apprised of his test performance, Bert's reaction was, "I'm not that good; those tests don't tell what you can do"; but he was obviously pleased. Reassurance and encouragement were needed over a three-year period of counseling with Bert before he was able to integrate into his self-picture the fact that he could probably succeed in college. His low record of marks caused him some concern when he was discussing college possibilities. He stated that he knew he could get by in high school without pushing himself. Bert had learned that he could, with a little cramming at examination time, pass courses in which he had low marks. He frequently held off in his efforts until final

examination time. By the end of his fourth year of high school, Bert had his sights upon college training, and although he turned down a scholarship in favor of enlistment, he had determined to begin a collegiate career after he was released from service.

Still another problem in test interpretation is indicated in the case of Bob who suffered from cerebral palsy. Some interesting speculations about his test performance and the efforts of a counselor to appraise his chances of success in future training are raised in this case.

BOB

Bob was highly regarded by both his teachers and his fellow students. He sought the counselor's aid in determining what he might be qualified for after graduation from high school. Bob's particular interests and leisure-time activities were in the field of railroad and airplane scale model building, electronics, and television production. He was particularly interested in exploring his chances of successful employment in these fields.

As early as the seventh grade one of his teachers remarked: "Bob will never do justice to himself on a test. He does not work well under pressure, and it takes him longer than normal to write. During the school year he did the regular work of the 7th grade, although his test record did not show this." At that time Bob's Stanford Achievement battery scores indicated a median grade placement level of 6.8 and his subtest grade placement scores ranged from 4.8 to 8.5. The higher scores were achieved in science and arithmetic. He had been placed in an orthopedic school from kindergarten on through his school career, and the following rather comprehensive set of test scores appeared on his records below.

His high school academic records showed a total of 8 A's; 10 B's; 10 C's; and one D. His highest marks were in English and social studies, while his lowest were achieved in mathematics and science.

Grade	Test	IQ Scores	Percentiles
Kinder- garten	Pinter-Cunningham	99	
4	Kuhlman-Anderson	96	
7	California Mental Maturity	107	
7	Leiter International Performance Scale	115	
7	Porteus Mazes	125	
8	Wechsler-Bellevue Scale I Verbal Scale	101	
11	California Mental Maturity	111	
12	Differential Aptitude Tests		
	Verbal Reasoning		07
	Abstract Reasoning		24
	Space Relations		34
	Mechanical Reasoning		15

A counselor's summary of Bob's high school career, written at the time a referral was made to a vocational rehabilitation agency, indicated that most of his vocational counseling had been centered around some phase of radio and television. Bob's major expressed interests were in the production and technical end of this work. His hobby of scale model building, electrical wiring, and electronics had been continued during his high school years. He had used these skills by doing most of the wiring and electrical work for the school plays. His work experience outside of school had been as a pinboy in a bowling alley, a bag-filler in a candy factory, and as a clerk and salesman in a hobby shop. He preferred the latter and expressed some interest in a job in that phase of retail trade during his junior year.

In his senior year, Bob's choices seemed to crystallize in the direction of technical electronics. He also explored the idea of form-maker or scale model builder in metal work or some phase

of the automotive industry. He realized that he could not compete on the open market in such demanding occupations because of the strain and tensions that would be present. He said that he felt able to do such work if he did not have to work under tension or at a rapid rate.

The counselor observed Bob under testing conditions several times, and felt that the conditions were such that no valid result could be expected. Bob had to strain to make the small marks needed on the machine-scored answer sheet, and he was unable to make his pencil do what he wanted it to do. His physical handicap was so severe that he could not type or write with enough speed or legibility to pass even beginning English class in college. Although he explored the possibilities of having an electronic aid to help him, or hiring someone to write for him, he seemed to be a poor risk for the usual college training.

Bob accepted his physical handicap very well. His interest in electricity, science, model building, and radio had all been started as therapeutic measures to enable him to gain control over his hands and the muscles of his fingers. The counselor's task was to determine whether these interests were to be regarded as simply healthy compensatory measures or potentially realistic vocational possibilities. In this case the test record could not be taken at face value. The dual handicaps imposed by the speed required, plus his inability to make the proper little marks on paper, could make all the results meaningless. Powerful motivation had evidently enabled Bob to accomplish, in his wiring and radio activities, that which would normally have been judged impossible. The resolution of this dilemma required serious examination of some of the assumptions that are commonly made about test performances.

In the following brief reports some other factors in the interpretation of test scores are raised.

DAVE

Dave achieved a perfect raw score of 90 for the 90 items he attempted on the Henmon-Nelson Test of Mental Ability when

he was a sophomore in high school. He ranked number one out of some 30,000 sophomores in a state-wide testing program. His school work was near perfect, and his teachers consistently awarded him A's. Dave wrote with a mature style, and his autobiography prepared for one of his counselors was insightful and well written. He had tentatively chosen some field of writing as his vocational goal. Yet Dave received a 15th percentile score on the Differential Aptitude Spelling subtest. Explanation of this one low score along with other scores on many tests, all at the 99th percentile, was hard to make. Dave has since graduated from a major university. The only exception to his straight A record was a B in one science course and 4 C's in ROTC.

MACK

Mack was a veteran and a junior in a school of education at a large midwestern university. He had been a building contractor and had planned, built, and sold more than a dozen houses before he entered college. It was difficult to interpret his 10th percentile score on the Bennett Mechanical Comprehension Test to him.

RAOUL

Raoul, an Indian boy from a tiny Spanish-American village in the Southwest, achieved a scaled score of 17 on the Block subtest of the Wechsler Intelligence Scale for Children. He was three years retarded in grade school because he did not seem to understand his lessons or even see why he was in school. His teachers were unanimous in their opinions that he was "stupid." Could a "stupid" person achieve such a high score in a complex mental function? What could a counselor tell Raoul or his teachers about his test performance that would be meaningful?

ART

Art, a sophomore in a premedical curriculum in a college in the Southwest, scored in the lowest quarter on the American Council of Education Psychological Examination and the Co-

öperative English test. His grade average was a high B, and he had received A's in chemistry and college algebra. During counseling he revealed very strong hostility feelings toward persons in authority positions and mentioned to the counselor that the college psychometrician, a former WAC major, made him so angry that his testing experience was not typical of his usual intellectual performance.

BILL

Deep-seated feelings of personal inadequacy and lack of personal worth bothered Bill so that he sought counseling. He wished for help in assessing his vocational choice of elementary school teaching, and wondered if he was intelligent enough to achieve his goal. His test scores were all in the top 10 percent for college freshmen, and his grades were all A and B, but he could not seem to accept these as evidence of his capabilities. He seemed unimpressed with the report of his test performance and unable to integrate it into his picture of himself, which had for many years been that of a person of little worth and little ability. The task of test score interpretation with Bill necessitated extensive counseling aimed at helping him rebuild new attitudes toward himself and the world about him.

BARBARA

Barbara, a high school junior, was a straight A student who scored above the 85th percentile on all subtests of the Differential Aptitude Tests and above the 90th percentile on the subtests of the Coöperative English test. Her vocational choice was clerk-typist. Her mother had held such positions. The realistic probability of immediate employment after high school as a secretary or typist seemed to outweigh the counselors' suggestions that her test performance and academic record suggested many more rewarding possibilities. She seemed totally unimpressed with her above average test scores and her fine academic record. Her inability to accept the implications of her above

average performance seemed to block efforts of the counselor to have her consider other employment opportunities.

The nine cases described above raise the issue of how many confirming data are needed to validate the assumption that test performances could provide adequate samples of behavior of each of the counselees described. In each case the test data would make a counselor hesitant about making any prediction about future performances of the individual, and the question must be raised about the need to interpret test scores in terms of other information about the counselee. Many of the current difficulties would seem to arise from the fact that counselors are trying to make individual interpretations of test performances when the counselee has taken a group test administered "by the numbers" and his score is compared with norms based upon what the *average* person in a group will do. When the individual case is examined, variability from rather than conformity to the general rule seems to hold.

Each of these cases illustrates one or more important considerations that must be noted prior to using test data for counseling. In the case of Bob, his test scores must be interpreted in the light of medical and psychological evaluations that are needed to bring meaning to his attempts to take tests to answer his questions about himself. Ed, Mack, or Art are examples of counselees with lower than average test performance that did not seem to be consistent with their academic performances. If one were to take the test performance at face value, great disservice would have been done to any one of these students. Each, in his own way, had achieved far greater goals than the test record would normally indicate. In each case the counselor needed to integrate the test data with other information that was contradictory. In Ed's case the counselor needed to assess the pressure to achieve academic goals raised by the family, and to speculate about the meaning of the grades received. Were they reflections of Ed's middle-class background, his compliancy, and his willingness to please his teachers? Or did they

truly represent achievement? Further investigation suggests that some of his poor test performances were caused by the tensions aroused by the family pressures to succeed. His counselor might conclude that the usual timed tests did not have personal validity for this boy.

Bert and Raoul, each in his own way, gave evidence of far more talent than his teachers had estimated. Part of this low evaluation of school achievement may be laid to their lower socioeconomic class backgrounds. In Bert's case this was strengthened by his manifest hostility to the social role of his teachers and to his refusal "to play the game" according to their rules as Ed had so well learned to do. Barbara, another student from lower than middle-class family, showed this in a different way. Although she had learned to achieve the academic standards set by the school, her refusal to seek the usual middle-class goals of upward social mobility through further education indicates that her personal value structure was little changed by contact with teachers. Her prime concern was to gain employment immediately upon graduation from high school despite the exhortations of teachers that she go on to college.

In Bill's case the test data were meaningless. Any prediction of his future success in training must take into consideration his feelings about himself. His counselor felt that until he could gain some insight into the reasons why he had adopted a self-concept as an inadequate, worthless person, he would experience difficulty in achieving success in anything he attempted.

"THAT DEPENDS"

The reader may have looked in vain in the discussion above for some formula or some scheme of weights that he could apply in the interpretation of an individual's test performance in relation to other data. The cry in guidance circles has been to be "objective."

This has caused counselors to distrust anything that does not have numbers attached to it. The writers feel certain that presentation of a series of Beta weights obtained by multiple regression equations would be more warmly received by many counselors than the more subjective, more clinical approach, ". . . that depends . . .," proposed here. Careful evaluation of the results of quantitative prediction for the individual case demands that counselors postpone its use for perhaps the next 50 years. Both the counselor and the counselee are apt to be deceived by pretensions to pure scientific objectivity when using the group measurement tools presently available.

The more "subjective" approach that has been proposed above demands counselors with high levels of training, persons sensitized to intrapersonal relations, and scientists with feelings of humility for their new and still unfinished techniques. The predictive equation may begin to balance when the counselor's appraisal of objective data about counselees is modified in terms of his experiences with like individuals. This flexible approach must be kept open for the interplay of insight, "feel," and intuition based upon carefully evaluated previous clinical experience and training. The questions "By how much?" "To what extent?" and "To what degree?" can be answered in part by the counselor's overall knowledge of the counselee. His knowledge of what the counselee has achieved in past performances and his assessment of the environmental pressures imposed upon the counselee's performances may contribute to the overall Gestalt of the counselor's diagnosis of a particular case.³²

At some future time the data involved in the ". . . that depends . . ." may be quantified. Until that time, the counselor should seek to exploit to the utmost its qualitative potentials. In doing so he may help the individual student to find his way and to

³² See: Carl R. Rogers. "Persons or Science? A Philosophical Question." *American Psychologist*, July, 1955, 10:267-278. Paul E. Meehl. *Clinical Versus Statistical Prediction*. Minneapolis: University of Minnesota Press, 1954.

estimate his best bets and best choices among the many avenues of action open to him in a free society.

INDIVIDUALIZED TEST INTERPRETATION

Throughout this volume, and particularly in this chapter, it has been suggested that test interpretation must be done in relation to other influential factors in the life history of students. Procedures for incorporating various items of background data in test interpretation have been discussed above, and some suggestions of how these might influence test scores have been made. It has been indicated that many personal background variables must be considered along with actual test scores when a counselor reaches the point in the counseling process at which test information will be of value.

The ways in which test data can be interpreted with other data and the manner in which interpretation can be done are controversial subjects.³³ The reader may consider what is involved when a counselor tries to use testing in his counseling. It may be assumed that he has used tests in answer to rather specific questions of his counselees instead of giving *all* available tests to *all* the students in the school in the hope that somehow they may prove useful at some later time. If this has been done, the counselor may assume higher motivation on the part of the counselee and, if this condition holds, the use of test results in counseling becomes a natural part of the counseling process.

Students in high schools are frequently sent to a large room and

³³ See: Robert C. Woellner. "Interpretation of Test Results in Counseling." *School Review*, December, 1951, 59:515-517. John W. M. Rothney. "Interpreting Test Scores to Counselees." *Personnel and Guidance Journal*, February, 1952, 30:320-322. Barbara A. Kirk. "Individualizing of Test Interpretation." *Personnel and Guidance Journal*, April, 1952, 30:500-505. Miriam Faries. "A Therapeutic Approach to Test Interpretation." *Personnel and Guidance Journal*, April, 1957, 35:523-526. Donald E. Super. "Testing and Using Test Results in Counseling." *Occupations*, November, 1950, 29:95-97. Edward S. Bordin. *Psychological Counseling*. New York: Appleton-Century-Crofts, 1955, pp. 273-279. Leona Tyler. *The Work of the Counselor*. New York: Appleton-Century-Crofts, 1953, pp. 142-166.

given one or more tests without their prior knowledge, consent, or interest and the outcomes of this procedure are apt to be disappointing for all concerned. Very little motivation is aroused by this routine, and if resistance and irritation are avoided, indifference and boredom seldom are. In such mass testing procedures the one most concerned with the outcomes, the student, is seldom informed of the results.

When tests are used as instruments in attempts to find partial answers to questions raised by the counselee himself, far more rapport may be assumed and a more valid set of scores expected.³⁴ It seems that students who have taken tests under such conditions should be apprised of the outcomes of their test performances. There is much opinion but little research on the value of doing so. One research study has, however, pointed up the fact that when high school students were given the results of their test scores during individual counseling sessions, their counselors observed no significant negative or disturbing effects.³⁵ Most students responded to this procedure with enthusiasm and appreciation. They indicated that they were thankful for help in establishing a more realistic picture of themselves through the counselor's aid.

Interpretation of test scores to counselees may provide partial answers to questions that have been raised in the counseling itself. The counselee may have asked, ". . . have I enough background in mathematics to consider a career in science?" The counselor's report of a test score in this area may provide a partial estimate. Test scores may also assist a counselee to see his test performance as a reflection of his expressed feelings about himself. Full consideration of this point is beyond the scope of this book but the reader is referred to reports by Bordin³⁶ and Tyler,³⁷ who have presented

³⁴ See: Ralph Bixler and Virginia Bixler, "Test Interpretation in Vocational Counseling," *Educational and Psychological Measurement*, Winter, 1946, 6:145-155. Paul E. Dressel and R. W. Matteson, "Effective Client Participation in Test Interpretation," *Educational and Psychological Measurement*, Autumn, 1952, 4:693-706.

³⁵ Rothney, *op. cit.*, p. 322.

³⁶ Bordin, *op. cit.*, pp. 274-276.

³⁷ Tyler, *op. cit.*, pp. 159-166.

Careful discussions of this matter. It is sufficient to say that valuable clues to the counselee's perceptions of self and his attitudes toward himself may be gained by discussion of his expected and actual test performances. The clues and hunches obtained may help in the attainment of fuller understanding of and by the counselee.

Since clear communication between counselor and counselee is essential, extreme care must be taken to use language that has real meaning to both. Technical terms such as IQ are subject to distortion or misunderstanding and should never be used. Generalized statements that avoid the inference the student has or has not ability to achieve at specific levels should be employed.

Test interpretations must be made in individual conference with the student. If attempts at interpretation are made in group sessions by teachers or others, there is much danger of misinterpretation and distortion by the student who is anxious about what his test results will foretell of his future.

During individual conferences the counselor should refresh the student's memory about the test he took by showing him a copy of the test. He may then indicate the student's percentile score and show how it compares with other students who have entered the kind of training the student is considering. The student may be told that in comparison with similar students his score is such that a given percentage of students scored as high as or higher than he did. If predictive implications are to be considered, the expectancy table described earlier may be used.

Insistence that test interpretations be made by trained counselors in individual conferences with students may emphasize the fact that test scores have limited meaning and that other information is needed before the questions that a student raises can be answered. This individual test interpretation procedure merges testing into the whole counseling process. It emphasizes development by the student of greater maturity and insight rather than upon the use of an added device that purports to give quick and final answers to

current problems. When thus relegated to their proper place, tests *can* be helpful in the counseling process.

SUMMARY

In this chapter some issues involved in interpreting test scores to counselees have been raised. Particular attention has been given to the way nontest data may be used with test scores in interpretations. The difficulties of basing predictions upon test scores alone have been explored and the counselor has been reminded of the various sources of error in such estimates. The effect of nontest variables upon prediction processes was explored, and special emphasis was given to the estimation of the influence of school marks, social class, position of the student's family, and various intrapersonal factors upon the total counseling process. Several cases were presented to illustrate the interaction of these forces in counseling. Emphasis was placed upon the necessity of understanding all aspects of a counselee's behavior. In the following chapter, discussion will focus upon attempts to measure personality, attitudes, and interests by use of structural and projective methods.

EXERCISES

1. *Develop in a role-playing session (with various students acting the part of the counselor, the parent, and the counselee) test score interpretation for George, a high school senior, who is considering entrance to a school of engineering. Assume in the first session that the following test scores are all you have to use. No other data about George are available.*
2. *Reconsider your role-playing interview with George and his parents with the additional data supplied by the following semester school marks earned in high school.*
3. *Again reconsider the role-playing interview with George with the addition of the following information taken from the counselor's files.*

careful discussions of this matter. It is sufficient to say that valuable clues to the counselee's perceptions of self and his attitudes toward himself may be gained by discussion of his expected and actual test performances. The clues and hunches obtained may help in the attainment of fuller understanding of and by the counselee.

Since clear communication between counselor and counselee is essential, extreme care must be taken to use language that has real meaning to both. Technical terms such as IQ are subject to distortion or misunderstanding and should never be used. Generalized statements that avoid the inference the student has or has not ability to achieve at specific levels should be employed.

Test interpretations must be made in individual conference with the student. If attempts at interpretation are made in group sessions by teachers or others, there is much danger of misinterpretation and distortion by the student who is anxious about what his test results will foretell of his future.

During individual conferences the counselor should refresh the student's memory about the test he took by showing him a copy of the test. He may then indicate the student's percentile score and show how it compares with other students who have entered the kind of training the student is considering. The student may be told that in comparison with similar students his score is such that a given percentage of students scored as high as or higher than he did. If predictive implications are to be considered, the expectancy table described earlier may be used.

Insistence that test interpretations be made by trained counselors in individual conferences with students may emphasize the fact that test scores have limited meaning and that other information is needed before the questions that a student raises can be answered. This individual test interpretation procedure merges testing into the whole counseling process. It emphasizes development by the student of greater maturity and insight rather than upon the use of an added device that purports to give quick and final answers to

current problems. When thus relegated to their proper place, tests *can* be helpful in the counseling process.

SUMMARY

In this chapter some issues involved in interpreting test scores to counselees have been raised. Particular attention has been given to the way *nontest data may be used with test scores in interpretations*. The difficulties of basing predictions upon test scores alone have been explored and the counselor has been reminded of the various *sources of error in such estimates*. The effect of nontest variables upon prediction processes was explored, and special emphasis was given to the estimation of the influence of school marks, social class position of the student's family, and various intrapersonal factors upon the total counseling process. Several cases were presented to illustrate the interaction of these forces in counseling. Emphasis was placed upon the necessity of understanding all aspects of a counselee's behavior. In the following chapter, discussion will focus upon attempts to measure personality, attitudes, and interests by use of structural and projective methods.

EXERCISES

1. Develop in a role-playing session (with various students acting the part of the counselor, the parent, and the counselee) test score interpretation for George, a high school senior, who is considering entrance to a school of engineering. Assume in the first session that the following test scores are all you have to use. No other data about George are available.
2. Reconsider your role-playing interview with George and his parents with the additional data supplied by the following semester school marks earned in high school.
3. Again reconsider the role-playing interview with George with the addition of the following information taken from the counselor's files.

TABLE 22. Test Scores and Norms of a Counslee

Tests	Grade	Score	Norms
Differential Aptitude— Numerical Ability	12	95th percentile	12th boys
Differential Aptitude— Mechanical Reasoning	12	45th percentile	12th boys
Differential Aptitude— Space Relations	12	30th percentile	12th boys
ACE Psychological— College Level	12		4-year college
Q Score		90th percentile	freshmen
L Score		70th percentile	
T Score		80th percentile	
Iowa Silent Reading	10	40th percentile	local norms
Otis			
Gamma	10	Otis IQ 108	national norms
Stanford Achievement, Advanced (J)	8	Battery Median	national norms
		8.4	
Paragraph Meaning		7.0	
Word Meaning		8.4	
Spelling		8.0	
Language		9.4	
Arithmetic Reasoning		9.0	
Arithmetic			
Computation		10.0	
Social Studies		8.2	
Science		8.5	
Study Skills		7.0	

Family Background: Father, newspaper publisher of country weekly, 8th grade education. Mother, college graduate, housewife. Older sister, Junior in College of Education.

Health: Slim, but tall for age. Frequent attacks of asthma as a child. Persistent trouble with sprained ankles in high school athletics. General health excellent.

Interests and Activities: Dates with girls at every opportunity. President of many school clubs, DeMolay, Hi-Y.

TABLE 23. Actual Grades of Student of Table 22

Rank in Class at Graduation	15th in a Class of 120
English	4 A's, 4 B's
Mathematics—Algebra	A, B
Geometry	C, D
Trigonometry	C
Language—Spanish	F, D
Social Studies—Civics	A, B
World History	B, B
U. S. History	C, B
Problems of Democracy	A, A
Science—General Science	C, C
Biology	C, B
Chemistry	D, C
Physics	C, C
Journalism	A
Speech	B
General Business	A
Typing	B
Band	5 A's, 3 B's

Collects match boxes. Likes dancing, tennis, badminton, travel.

Work Experience: *Works for father in press room (a job he dislikes). No other work.*

Problem Areas: *Father insistent on engineering because of high status and high pay. Boy neutral or passive toward this selection. When asked for alternate choice, George was unable to say other than "working with people in some way."*

4. Read the following case of Mike. "I believe the one reason why my son Mike is attending the University," wrote his mother to a school official, "is because at one time you told him he had the ability, but it was up to him to use it. That was all the encouragement he needed. After graduation he attended summer school and took a mathematics course which he knew he didn't know very well and finished with a 92 average. Besides he worked and saved enough money to pay his entire tuition and expenses. He is doing

very well at the University. I want to compliment you on your fine work."

No one, at the time Mike was graduated with a rank of 186 in a class of 353, would have predicted that he would go on to a university. In fact, one month before he was graduated Mike said that he expected to go to work in a factory or as a handyman in a hotel. He said that if things went well with him he would be "just lucky." Yet Mike surprised everyone, even his mother, and certainly his teachers. He was one of those persons who make prediction of human behavior a very hazardous procedure.

SCHOOL RECORD				
Subject	Grades			
	9	10	11	12
English	C B	C D	B C	C C
Spanish			F D	
Amer. History			C B	
Civics	B B			
Geography				B B
Amer. Problems				A A
Algebra		D D		
Biology		D D		
General Science	B B			
Chemistry			B B	
Physical Ed.	Credit	Credit	Credit	Credit
Industrial Arts		C D		
Agriculture	B C			

Mike described some of the highlights of his career in the following autobiography, written for his eleventh grade English teacher. His own story, just as he wrote it, is reproduced below.

Oct. 15, 1932 was an important date to me because that was the day I was born in a city hospital.

I have never known my father as he died in an automobile wreck before I was six weeks old. I have only known one relative on my fathers side, my uncle one time congressman for the state. I have also been told that my Grandfather was once govenor of a state.

My Mother ran a shop somewhere in the city. I remember little of this but two things have stood out quite clearly, one was that I had gotten in to a fight with another boy about my age. This boy pick up a board which had a nail in it and the nail price my eye. Everything turn out alright as the nail had missed the pupil. The only other thing I remember of those yrs. was the time I had a tooth pull.

During the depression my mother sold the shop got work as a housekeeper.

I had now started to school.

Well It didn't take look when I found a friend in this school and we went together in to stores and started to lift candy. We were soon caught & I probably was never so scared in my life.

There was a dog in my life a that time teddy was his name he was old & blind but was dreadfully smart.

We then moved to a new house and mother took in my Grandmother as she was very ill. I remnber brief but Pleasant visits to my grandfather's in another town. For a summer or two we live a lake five, a small wooded lake part of which is now a game refuge.

When I reached the age of eleven mother married so we move to another town.

I supposed I fooled around like the average boy the only important thing was that the war was soon over with. You see I had remenber Pearl harbor and I remenber mother reading the papers to me when hilter had first started waring on neigrning countrys in 38 or 39.

I had used to sat near the window & watch what was than New car as the went down the street as she read.

A little while after the Pres. death we had brought a farm

near here were I now live. I have travel quite bit these last few years too. That about all the important things.

Despite Mike's lack of polish in his written work he received average marks in English during his high school career. The English teacher in his junior year said: "His spelling is barbarous, but what he says is really bright. He has unusually fine taste."

Mike had always been required to work hard at part-time jobs. He did farm work, clerked in a store, and, during his senior year, worked as a houseman at a neighboring city hotel three or four evenings a week and all day on Saturdays and Sundays. Since, on the evenings he worked, he did not get home until one o'clock in the morning, he was often sleepy in class and seldom got his homework done. As the result of an accident, in which he suffered some concussion, Mike had frequent headaches. This condition and the lack of sleep resulted in the description of him by his teachers as the "epitome of apathy and lethargy."

Mike did not seem disturbed when he was told what he had done on the tests. He did say that he had one of his headaches when he took the Primary Mental Abilities Test, and that he always expected to do poorly in any form of mathematics. He liked algebra least of all the subjects he took in high school and Spanish was not interesting because he could not get good grades. English was one of the subjects he liked best because, he said, "I like to read and write stories." Biology was also a favorite subject. Because at one time he thought he might go into defense work after graduation, Mike added to his school and work load by taking an evening course in welding at the city vocational school two nights a week. The additional time meant even less rest than he had been accustomed to, and he became even sleepier and more lethargic during his senior year.

His test record follows below.

Mike was "on his own" as far as activities were concerned. Early interests in chemistry and stamp collecting soon vanished when he raised enough money to buy a cheap, old car. He liked to sing and play baseball, but he took no part in school activities

Tests	Percentile		
	Grade 10	Grade 11	Grade 12
Henmon-Nelson Test of Mental Ability	58	75	
Reading Tests			
<i>Progressive</i> Reading Vocabulary		X	
<i>Progressive</i> Reading Comprehension		T	
Coöperative Speed of Reading			
Primary Mental Abilities			
Verbal	50		
Space	32		
Reasoning	15		
Number	9		
Word Fluency	12		
Differential Aptitude Tests			
Number Reasoning		5	
Sentences			50
Spelling			15
Mechanical Reasoning			40
Space Relations		15	
Clerical Speed and Accuracy		40	
X 1½ years accelerated			
T 2 years accelerated			

in those areas because, he said, "I guess I'm a little lazy on those things." He liked to read a lot of "any kind of books except cheap novels." His best experience was a two-week trip with his uncle to Chicago where he enjoyed seeing things, "especially skid row."

In a check list of activities, he marked the following items in the frequency given.

Reading	}	Every day
Radio		
Movies	}	Once a week
Television		
Tennis		
Hiking	}	Twice a week

He said that he could go out any evening that he wished and could come home at any time that he chose on any night of the week. He refused to fill out a diary of activities illustrating a typical week but he said he was seldom at home, and he implied that no one was much concerned whether he was there or not.

Mike did not remember his father. His stepfather merely accepted him, but neither showed much concern about the other. His half-brother and half-sister were nine and twelve years younger, and he was indifferent about their activities. His mother, who had had only an elementary school education and who had struggled hard to make her own way in depression years, was convinced that Mike should go on to training beyond high school. She saw no way of providing financial assistance, and he seemed completely reconciled to the idea that he would have to make his own way as he had done throughout his high school career. His stepfather, who had been graduated from high school and attended a vocational school, was a part-time farmer and carpenter who saw no reason, even if he had been able to do so, to support his stepson. Mike said that he never talked about his hopes and plans at home.

Any plans for Mike's future would be influenced by his appearance and manner. He was slow-moving and slow-talking. He had a serious case of acne. It seemed as though he had slept in his clothes and had not had time to comb his hair before coming to school. And this condition did not change significantly at the time when most of his contemporaries became concerned about their appearance. He seemed to be very sleepy at all times, and the

general impression he gave was one of a "lost soul" who seemed to lack confidence in himself and about whom nobody seemed to care. His two closest friends had notorious disciplinary records, and Mike said that they would describe him as "quiet in class but reckless otherwise." From observations of Mike in school, the term "reckless" would just never seem to apply.

When he was a sophomore, Mike said that he would like to join the Air Corps for three years and then become a farmer. He continued to express interest in the armed forces in his junior year because, he said, he would learn to "take and give orders and to learn to think in a clutch." When he was a senior he had changed his mind. At that time he indicated he would not enlist. "It's long enough when you're drafted," he said. During the first half of his senior year he seemed very uncertain about the future. He showed some interest in going to college to study accounting, and he wanted to take some tests to determine whether or not he had the aptitude for it. The pros and cons of several occupations were discussed when he said that he wanted to learn more about them. In the last month of his senior year, however, he said that he would take a job in a factory. No one predicted that Mike would set up a plan to work his way through college and to carry it through in the manner described by his mother in the first paragraph of this case study.

- a. Comment on Mike's two Henmon-Nelson scores, i.e., 58th percentile and 75th percentile. Why the disparity between them?
- b. What personal factors might account for the difficulties in prediction for the future with Mike?
- c. Suppose Mike had come to you as his counselor in his 12th year of school and asked your opinion about his chances for success in a university. What could you have told him?

REFERENCES

Berdie, Ralph. *After High School, What?* Minneapolis: University of Minnesota Press, 1954.

- Caplow, Theodore. *The Sociology of Work*. Minneapolis: University of Minnesota Press, 1954.
- Carter, Ralph S. "How Valid Are Marks Assigned by Teachers?" *Journal of Educational Psychology*, April, 1952, 43:218-228.
- Eells, Kenneth, and Others. *Intelligence and Cultural Differences: A Study of Cultural Learning and Problem Solving*. Chicago: University of Chicago Press, 1951.
- Guilford, J. P. *Fundamental Statistics in Psychology and Education*. New York: McGraw-Hill, 1956.
- Gustad, John W. "Test Information and Learning in the Counseling Process." *Educational and Psychological Measurement*, Winter, 1951, 11:788-795.
- Heimann, Robert A., and Schenk, Quentin. "Relations of Social Class and Sex Differences to High School Achievement." *School Review*, July, 1952, 49:213-221.
- Hollingshead, August. *Elmtown's Youth*. New York: Wiley, 1949.
- Hollingshead, Byron. *Who Should Go to College?* New York: Columbia University Press, 1952.
- Hull, Clark. *Aptitude Testing*. Yonkers, N.Y.: World Book, 1928.
- Leonard, W. N. "Psychological Tests and the Educational System." *School and Society*, April 12, 1952, 75:225-229.
- McCabe, George E. "How Substantial is a Substantial Validity Coefficient?" *Personnel and Guidance Journal*, February, 1956, 34:340-344.
- Meehl, Paul. *Clinical versus Statistical Prediction*. Minneapolis: University of Minnesota Press, 1954.
- Rothney, John W. M. "Evaluating and Reporting Pupil Progress." *What Research Says to the Teacher*, No. 7. Washington, D.C. National Education Association, 1955.
- Rothney, John W. M., and Roens, Bert A. *Counseling the Individual Student*. New York: Dryden Press, 1949.
- Smith, Eugene R., and Others. *Appraising and Recording Student Progress*. Adventure in American Education, Vol. 3. New York: Harper, 1942.
- Thorndike, Robert L. *Personnel Selection*. New York: Harper, 1949.

- Warner, W. Lloyd, Havighurst, Robert L., and Loeb, Martin B. *Who Shall be Educated?* New York: Harper, 1944.
- Wesman, Alexander. "Better Than Chance." *Test Service Bulletin*. New York: The Psychological Corporation, 1953, 12 pp.
- Wolfe, Dale L. *America's Resources of Specialized Talent*. Report of the Commission on Human Resources and Advance Training. New York: Harper, 1954.

CHAPTER VIII

Personality Questionnaires and Interest Inventories

The terms "inventories" and "questionnaires" are used in the title of this chapter to emphasize the fact that there are *no* interest or personality *tests* in any sense of that word. Although there are instruments that bear such titles¹ and the terms "personality" and "interest" tests are commonly used by educators, personnel workers, and psychologists, there is really no justification for their use. These instruments are either questionnaires such as the Kuder Preference Record and the Strong Vocational Interest Blank or controlled interviews such as the individual form of the Minnesota Multiphasic Inventory and the Rorschach.² As questionnaires and interviews they possess all the limitations as well as some of the advantages of such techniques but they should not be described or thought of as *tests*. It is essential that counselors recognize the difference between inventories and tests and that they stress this

¹ See, for example, *The California Test of Personality*, California Test Bureau. *The Thematic Apperception Test*, Harvard University Press. *The Rogers Test of Personality Adjustment*, The Psychological Corporation.

² There are more than one hundred of such instruments. See those listed in the several *Mental Measurements Yearbooks* by O. K. Buros, published by Rutgers University Press and Gryphon Press.

difference in their work with counselees, to prevent the common tendency to misinterpret the scores.

Counselees who have heard about "scientific" tests of aptitude insist on reading into questionnaires and inventory scores something that was never intended by most of their authors and for which there is absolutely no justification. They tend to think that they have taken a vocational fitness test or a measure of aptitude for some educational experience. Since these instruments use formats similar to tests, and since they have scores and norms, they encourage this kind of misinterpretation. It is most unfortunate that authors and publishers of inventories or personality questionnaires have not taken enough action to prevent the misuse and misunderstanding that are common among those who take them, and even among those who administer them.

There is a great deal of evidence that personality questionnaires, controlled interviews, and interest inventories are widely used in counseling.⁸ Just why this should be so in view of the demonstrated inadequacies of these devices is difficult to understand. It seems that it must be a combination of amazing, psychometric innocence on the part of the users, naïveté in considering the counseling job as a "quickie" affair rather than a complex longitudinal problem, mistaken faith in statistics on the part of inventory producers and consumers, expediency, and a desire to keep up with the other fellow who uses them for any of the above reasons. Perhaps another reason for their popularity can be found in the seeming exactness they give to the counselor's work. Counseling interviews may seem not scientifically respectable enough to impress one's colleagues or clients, but an array of scores seemingly supported by pedantic jargon might possibly do so. The popularity of the instruments may be due in part, then, to the psychological support that counselors, working in a relatively new area, and without adequate

⁸ See, for example, Ralph F. Berdie, "The State-Wide Testing Programs," *Personnel and Guidance Journal*, April, 1954, 32:454-459; and J. R. Berkshire and Others, "Test Preference in Guidance Centers," *Occupations*, March, 1948, 26:337-343.

CHAPTER VIII

Personality Questionnaires and Interest Inventories

The terms "inventories" and "questionnaires" are used in the title of this chapter to emphasize the fact that there are *no* interest or personality *tests* in any sense of that word. Although there are instruments that bear such titles¹ and the terms "personality" and "interest" tests are commonly used by educators, personnel workers, and psychologists, there is really no justification for their use. These instruments are either questionnaires such as the Kuder Preference Record and the Strong Vocational Interest Blank or controlled interviews such as the individual form of the Minnesota Multiphasic Inventory and the Rorschach.² As questionnaires and interviews they possess all the limitations as well as some of the advantages of such techniques but they should not be described or thought of as *tests*. It is essential that counselors recognize the difference between inventories and tests and that they stress this

¹ See, for example, *The California Test of Personality*, California Test Bureau, *The Thematic Apperception Test*, Harvard University Press, *The Rogers Test of Personality Adjustment*, The Psychological Corporation.

² There are more than one hundred of such instruments. See those listed in the several *Mental Measurements Yearbooks* by O. K. Buros, published by Rutgers University Press and Gryphon Press.

difference in their work with counselees, to prevent the common tendency to misinterpret the scores.

Counselees who have heard about "scientific" tests of aptitude insist on reading into questionnaires and inventory scores something that was never intended by most of their authors and for which there is absolutely no justification. They tend to think that they have taken a vocational fitness test or a measure of aptitude for some educational experience. Since these instruments use formats similar to tests, and since they have scores and norms, they encourage this kind of misinterpretation. It is most unfortunate that authors and publishers of inventories or personality questionnaires have not taken enough action to prevent the misuse and misunderstanding that are common among those who take them, and even among those who administer them.

There is a great deal of evidence that personality questionnaires, controlled interviews, and interest inventories are widely used in counseling.³ Just why this should be so in view of the demonstrated inadequacies of these devices is difficult to understand. It seems that it must be a combination of amazing, psychometric innocence on the part of the users, naïveté in considering the counseling job as a "quickie" affair rather than a complex longitudinal problem, mistaken faith in statistics on the part of inventory producers and consumers, expediency, and a desire to keep up with the other fellow who uses them for any of the above reasons. Perhaps another reason for their popularity can be found in the seeming exactness they give to the counselor's work. Counseling interviews may seem not scientifically respectable enough to impress one's colleagues or clients, but an array of scores seemingly supported by pedantic jargon might possibly do so. The popularity of the instruments may be due in part, then, to the psychological support that counselors, working in a relatively new area, and without adequate

³ See, for example, Ralph F. Berdie, "The State-Wide Testing Programs," *Personnel and Guidance Journal*, April, 1954, 32:454-459; and J. R. Berkshire and Others, "Test Preference in Guidance Centers," *Occupations*, March, 1948, 26:337-343.

evidence of their effectiveness, may feel that they need; and the round-the-clock hucksterism in the sales of the instruments must account in large measure for their widespread use. Certainly it cannot be justified on the basis of logical reasoning or experimental evidence.

LIMITATIONS OF SHORT-CUT METHODS

Various writers have demonstrated that all the instruments for measuring personality and interest have such serious limitations that their use seems likely to do more harm than good. Among those limitations the following have been noted.

1. Scores can be faked deliberately or without awareness.
2. Titles of the instruments are simply christenings by their authors.
3. The vocabulary used is a source of confusion to subjects who take the instruments.
4. Many of the instruments force the subjects to make choices among items about which they have neither knowledge nor concern. They may also require choices among items of unequal familiarity.
5. Use of particular scoring methods in order to get so-called "objectivity" limits the subjects' expression of enthusiasm or concern.
6. Statistical methods used in construction and norming of the instruments are questionable.
7. Results are subject to misinterpretation by those who take the instruments.
8. Evidence of the predictive validity of the instruments is either nonexistent or questionable.
9. They rely on self-estimates known to be highly invalid.
10. The cultural background of the subject is not given adequate consideration.
11. They suggest stability of personality and hence encourage-

ment of counseling of subjects as they are without consideration of what they may become.

12. In extreme cases of high interest or disturbed personality they may simply elaborate the obvious.

13. They discourage experimentation because they seem to provide a large quantity of numerical scores for rapid calculation.

ATTEMPTS AT JUSTIFICATION OF SHORT-CUT METHODS

Before consideration is given to each of the limitations listed above it should be noted that *interest and personality testers rely frequently on a superficial kind of reasoning that sometimes seems to justify the use of their instruments. It usually runs something like this: "If you are going to be successful in any undertaking you must have interest in it or the personality for it, as well as ability to do the required tasks. It will be necessary then to inventory your interests and personal traits to see if you have them. If your inventory and ability scores are high then you are likely to succeed in the tasks you are about to undertake."* From this kind of statement they then go on to administer the instruments and counsel on the basis of the results. Typical of such statements is the following.

The fundamental reason why general clinical counselors and researchers in personnel work have devoted increasing attention to "interest" lies in widespread common observations of workers in different fields. Such observation reveals *simply* that, granted the presence of abilities commensurate with the demands of the job, workers are successful, happy in, and satisfied with their jobs if they feel at home and at ease with those with whom they have to work. This feeling of at-homeness arises from having strong and large areas of common interests which spread far beyond the borders of on-the-job behavior. . . . If, as observers, we move *suddenly* from association with preachers to hobnob, say with horse-racing stablemen, liquor salesmen, or ballet dancers, we are struck at once by the marked difference in the interests of each group. We may readily and

quite accurately conclude that a successful worker in one of these occupational sets would feel like a fish out of water if he attempted to carry on the job of another. A preacher, for example, might have all the fine build, muscular coördination, sense of rhythm, and timing of a potentially great ballet dancer, but his interest would wholly inhibit him from trying to be or succeeding in becoming one.* (*Italics added.*)

From this kind of illustration in which extremes of interests in occupations *not* covered by interest inventories are used for illustration, and in which such impossible situations as comparing a person's feelings to a "fish out of water" are used, the authors in this typical comment jump to the clincher in these words: "One of the most important functions of the counselor in educational or industrial practice, therefore, is helping of individuals to match their aptitude and ability patterns with their interest patterns."

From such statements comes the attempted justification of having students respond to hundreds of items by throwing a circle around an L, an I, or a D to indicate liking of, indifference to, or dislike for such items as the names of occupations, school subjects, amusements, activities, peculiarities of people, listing of preference for activities, comparisons of two activities, and self-rating of characteristics. The scores are then summed and the total is interpreted to indicate the similarity of interest of the client with those who have been successful in the occupations indicated, but *not* for preachers, horse-racing stablemen, liquor salesmen, or ballet dancers. Tyler³ and Hahn and MacLean⁴ have pointed out that one of the most commonly used instruments, *The Strong Vocational Interest Blank*, is not appropriate for investigating interests of youth much below the age of 17, and is not as useful in the high school as in the college situation. Its range of occupational keys outside the profes-

* Milton E. Hahn and M. S. MacLean, *Counseling Psychology*. New York: McGraw-Hill Book Co., 1955, p. 196.

³ Leona E. Tyler, *The Work of the Counselor*. New York: Appleton-Century-Crofts, 1953, p. 132.

⁴ Milton E. Hahn and Malcolm S. MacLean, *Counseling Psychology*. New York: McGraw-Hill Book Co., 1955, p. 207.

sional fields is so limited that it is not informative for the majority of students who will not go into professions.

The long jump from the generalizations from "common observation" to elaborate profiles indicating various degrees of so-called vocational interest or personality adjustment is made quickly. It is implied, despite the limitations noted above, that the inventory or questionnaire makes the jump possible and that counseling can be based in part upon them. Although careful writers surround their statements about the value of such devices with cautions about their exclusive use many still cannot resist such statements as these.

If pressure of time forces us to use a single method, most counselors prefer to depend upon a well-standardized measurement because of its greater demonstrated reliability and validity.⁷

It may be generalized that some caution is required in basing vocational decision upon interest test results during the early high school years, and that repeated measurements of interests is desirable for adequate guidance procedures at this age level.⁸

Anyone who is in a position to take a personality inventory can, if he so desires, answer its questions honestly and thus get a reasonably adequate basis for personality improvement.⁹

In the case of personality traits or mental health, use has been made of the judgments of individuals who are intimately acquainted with the subject being tested. However, the responsibility for such judgments has been such as to render them well-nigh useless . . . carefully constructed personality inventories thus are probably in some instances more valid than criteria with which they are correlated.¹⁰

Projective methods are peculiarly appropriate to schools because you can give them without exciting the subject and arousing emotional resistance or disturbance. . . . They can be used for some of

⁷ Hahn and MacLean. *Ibid.*, p. 210.

⁸ R. Jacobs. "Stability of Interests at the Secondary School Level." *Educational Records Bureau Bulletin*, July, 1949, No. 52, pp. 83-87.

⁹ L. P. Thorpe. *The Psychology of Mental Health*. New York: The Ronald Press Co., 1950, p. 252.

¹⁰ L. P. Thorpe. *Ibid.*, p. 648.

the perplexing disciplinary problems; namely to find kleptomaniacs, perverts, liars, those who write obscene notes or circulate obscene drawings because you can locate such individuals. They will reveal themselves in various ways, as in the T.A.T., for example. They can't help but tell what is going on inside, as has frequently been shown.¹¹

The high school or college teacher who is attempting to help his students in the choice of future occupations may find either or both of these tests (Kuder and Strong) useful.¹²

Clear-cut interest types seem to exist, and can be isolated on this test as early as the tenth grade. The interest types have definite personality correlates, somewhat indicative of the statement that birds of a feather flock together. Over a period of time there is considerable stability of measured interests.¹³

When an individual pupil is being counseled with regard to college preparation, if evidence of high aptitude in a certain field, let us say science, is supported by objective evidence of interest in the same area, the pupil may be advised with greater confidence to consider majoring in this area in college, for it is known that interest in a field increases the likelihood of success in that field.¹⁴

OBJECTIONS TO SHORT-CUT METHODS

Many workers in the field of education, psychology, and guidance have voiced opinions that are completely in contrast to those cited above. Samples of such statements follow.

It is certain that the matching of internal questionnaire factors and external behavior factors is systematically more difficult than person-

¹¹ Lawrence K. Frank. "Understanding the Individual Through Projective Techniques," in Arthur E. Traxler (Ed.), *Goals of American Education*. American Council on Education Studies, No. 40, Washington, D.C.: American Council on Education, April, 1950, p. 61.

¹² W. C. Morse and G. M. Wingo. *Psychology and Teaching*. New York: Scott-Foresman & Co., 1955, p. 402.

¹³ John G. Darley. "Tests and Personnel Work in College," *New Directions for Measurement and Guidance*, American Council on Education Studies No. 20. Washington, D.C.: American Council on Education, August, 1944.

¹⁴ Arthur E. Traxler. "Guidance Toward College Preparation." *School and Society*, February 25, 1931, 74:113-116.

nel workers imagine. . . . The self-inventory represents the nadir of scientific invention and subtlety.¹⁵

An overwhelming majority of publications presuming to be validity studies (of structured tests of personality) reported significant findings. It was difficult, however, to get a clear picture of the validity of the inventories. Frequently investigators ran a whole series of significant tests involving one or more inventory scales and, finding that one or two out of 20 differences or relationships reached an acceptable level of probability, claimed some validity for the inventory. What they apparently failed to realize is that such a series of significant tests is also subject to sampling considerations, and that the relatively few successful instances may have arisen purely by chance. In any case the need for cross-validation and replication is great.¹⁶

The scientifically minded research worker who reads a sample of the studies reported above [on projective techniques] is likely to feel at least mildly disturbed. He will want to point out that research workers with projective techniques often omit control groups or cases, leave out important descriptions of their subjects and methods, do not use enough subjects, employ faulty designs, pay little attention to norms, continue to employ concepts that research has shown to be faulty, interpret the product of "test plus user" effect as a product of the test itself, describe criteria vaguely, present tentative reports as if they were final, and often introduce irrelevant information into their studies.¹⁷

It would be helpful if interest inventories had built into them or, more realistically, if psychologists had available for coordinate use, measures which would differentiate the interest profile which adequately reflects an adequate self-concept from the interest profile

¹⁵ Raymond B. Cattell, *Description and Measurement of Personality*. Yonkers, N.Y.: World Book Co., 1946, pp. 341-342.

¹⁶ Edward J. Furst and Benno G. Fricke. "Development and Applications of Structured Tests of Personality." *Review of Educational Research*, February, 1955, 26:26-55, p. 27.

¹⁷ John W. M. Rothney and Robert A. Heimann. "Development and Applications of Projective Techniques." *Review of Educational Research*, February, 1956, 25:55-71, p. 66.

which is the result of inadequate perception either of self or the world of work.¹⁸

VALIDITY OF INTEREST INVENTORIES

Now that the reader has seen samples of contrasting *opinions* on the value of interest inventories, personality questionnaires, and structured or projective interviews for assessment of interest and personality, he may want to consider evidence about their utility. The counselor will look first at the evidence of validity and, for his purposes, particularly at data on predictive validity. It may come as a shock to find that, although the instruments are widely used in guidance, the evidence of validity is scant.

It is common practice to report as evidence of validity of interest inventories the finding that some scores, say of nurses, are slightly higher than others on social service scales, or that accountants score significantly higher on a computational scale than persons who are not in that occupation. The counselor will not be impressed with such data. He will tend to dismiss the slight relationships reported as elaborations of the obvious, since persons who do not like computation are unlikely to enter an occupation requiring it if they can resist pressures or know enough about it.¹⁹ He will note that, even if there are significant differences between groups in and out of an activity, there can be much overlap between their scores and that his particular counselee might be one of those cases in the area under the curves of distribution in which there is no discrimination between groups. He will look for evidence that inventory scores obtained by expenditure of much money and time will tell him

¹⁸ Donald E. Super, "The Measurement of Interests," *Journal of Counseling Psychology*, Fall, 1954, 1:171-179.

¹⁹ It has been found, for example, that about half of a sample of university students do not show clear-cut patterns of interests. See: John G. Darley, *Clinical Aspects and Interpretations of the Strong Vocational Interest Blank*, New York: The Psychological Corporation, 1941, pp. 19-20. See also: Edward A. Lincoln, "The Insignificance of Significant Differences" *Journal of Experimental Education*, March, 1936, 2; and Ralph W. Tyler, "What is Statistical Significance?" *Educational Research Bulletin*, March 4, 1931.

counselor of high school students or college freshmen. The interpretation of such data to counselees is an almost impossible task. If the counselee is told that persons who have completed such long and costly training for, say, accountancy, and have been successful in their work, tend to mark an inventory in much the same way that they did while they were in training, is he to conclude that the inventory is valid for high school students who say that they want to go into such occupations? Do not the findings just suggest that specialized training tends to produce less flexible persons? Does it not then seem possible that students may develop the suitable interest patterns by taking the specialized training? Does evidence that persons who score higher on the insurance salesman score than others indicate anything other than the fact that some good salesmen of insurance are also good sellers of themselves on the inventories? And, of course, since individuals are irreversible it cannot be shown that they might not, after all, have been better social studies teachers than insurance salesmen.

It is often argued that performances and choices of *recent* graduates provide no criteria of whether the students have chosen to enter fields of their interests. If this is so, authors of inventories and questionnaires should state this very clearly in their manuals, and counselors should pass the information on to their clients. The counselor's job will then become more complex. He will have to make it clear to counselees that these scores have little to do with the next few years but will be important several years later. Those who know high school students well will find the suggestion amusing. College students will, of course, be planning further ahead than random high school students.

It seems from the above that the validity of the inventory on which the most follow-up data are available is, to say the least, questionable. The manuals of the various inventories and the many short-term studies in the literature do not clear up the doubts that one must have about the validity of such instruments.²¹

²¹ See the numbers of the *Review of Educational Research* entitled "Educational

REASONS FOR LACK OF VALIDITY

One of the outstanding weaknesses of the inventories, apparent to anyone who will look and fully established by research,²² is the fact that their forecasting efficiency is severely limited by the faking of responses. To those who claim that a subject may not fake responses when good rapport has been achieved, it should be pointed out that there is no way to tell whether or not he has faked his answers. Every score must be questioned regardless of the conditions under which it has been obtained.

All faking may not, of course, be deliberate. The subject may not be aware that he is cooking the results while he is actually in the process of doing so. Consider, for example, the following case:

Jim, a sophomore in college, had become engaged to the daughter of a wealthy business man who had promised to set him up in the business after the marriage. Jim had some qualms about becoming dependent on his prospective father-in-law and decided to make a career for himself in psychology. At this period he made a high score for psychologist on a well-known interest-measuring device. After a year of psychology, during which his grades were low and the attraction of a good business opening was having less influence on his conscience, he decided to accept his father-in-law's offer. He took the interest-measuring device again, scored high this time in business and found in his score a satisfactory justification for his decision. After all, the tests showed that he was interested in business.

and *Psychological Testing*" appearing at various intervals from June, 1932, to February, 1956, and the *Annual Reviews of Psychology* for summaries of and bibliographic reference to such studies.

²² See: Arthur L. Benton and S. I. Kornhauser, "A Study of Score Faking on a Medical Interest Test." *Journal of the Association of American Medical Colleges*, February, 1948, 57-60. Edward S. Bordin, "A Theory of Vocational Interests as Dynamic Phenomena." *Educational and Psychological Measurement*, Spring, 1943, 3:49-63. Donald S. Patterson, "Vocational Interests in Selection." *Occupations*, May, 1946, 152-153. Verne Steward, "The Problem of Detecting Fudging on Vocational Tests." *Personnel Reports for Sales Executives*, January, 1947. H. P. Longstaff, "Fakability of the Strong Vocational Interest Blank and the Kuder Preference Record." *Journal of Applied Psychology*, August, 1948, 360-369.

Perhaps Jim's case is not typical and it might even be argued that the scores provided some useful hypothesis about him, but the inventory results seem only to have elaborated the obvious since he had faked it to suit his stated desires. The point here is that faking can occur and the counselor can never be quite sure about the absence or presence of deliberate or unconscious faking of responses. It is possible that results may be cooked for such reasons, among others, as attempts to justify acceptance or rejection of a stated choice, desire to please or disturb someone who will see the scores, attempts to beat the game for the pleasure of doing so, and wishful thinking. Suggestions by some authors for overcoming the faking of responses such as emphasizing speed, trying to encourage students to be honest, and viewing very high scores with suspicion would not be necessary if the instruments really did the job they are designed to do.

But assuming that there has been no deliberate or unconscious faking of responses, there must always be a question about the subject's reason for marking the items the way he does. The following anecdote told by a school counselor seems pertinent:

A semiliterate country boy, to whom the Kuder Preference Record had been administered, walked into my office, test profile sheet in hand. A glance at the percentiles revealed that his highest score was in the Literary area—around the 85th percentile. Not immediately taking up discussion of his test results, I talked with him a bit. As judged by the caliber of his responses, and the general level of his communicative ability, I marveled at the high "expression of literary interest." Finally I remarked upon his high score and asked him if he read a great deal. "Naw," he replied, "I don't read nawthin' much—onct in a whal a detectif magazine." When further pressed concerning his choice on the preference record, he explained that he thought it would be nice to be an author with nothing to do but write books.

Although there was no follow-up in this case, the other data

available on this boy indicated that a literary career for him was highly unlikely.

The reasons why a person indicates the preferences he does must always be subject to misinterpretations. To do counseling based on lists of responses without knowing the *why* of them is to perform at a very low level. Counselors, if they are to be effective, must work at a level beyond that which is implicit in securing rapid and superficial estimates of abilities, aptitude, and interests.

One of the most disturbing factors in the interest measurement movement is the number of counselors who do not realize that the names of some of the inventories or subsections of them, and the titles indicated by scores on their items, are simply christenings. An author frequently chooses to name a test, a group of items, or a single item as, say, a measure of mechanical interest because he thinks it is. He chooses to christen another item as a measure of clerical interest because it appears to him that that is what it might be. Then, of course, after he has christened it he chooses a system of attaching numbers to it, and it seems to become a scientific objective test. The internal consistency indexes that the test authors employ indicate only that they have been fairly consistent in the christening process. The reader might find it interesting to guess the area in which the following items, taken from inventories, are said to be measures of preference or interest and then look up the key to see how they are actually scored. The items are: build bird houses; doctor horses, cattle, or hogs; care for and repair people's teeth; develop a variety of pitless cherry; be an expert in cutting jewels; camp out; sleep in a tent or in the open; enter the bureau of printing and engraving; be a psychologist.

The surprises that one gets from such an exercise are many and the implications great. An author has christened an item in one way, but the counselor might christen it otherwise and would challenge the author's right to give it the name he has given. If the counselor were to continue the exercise and continue to challenge seriously the groupings of items and the names given to them, he

would then wonder if anyone could be expected to take the scores seriously.

The point about the christening of items does not negate, of course, the previous point about the faking of items. There will be enough obvious items about which most persons can agree at least at a superficial level. Some of the dumbest counselees will recognize that "playing a piano" is supposed to be an indication of musical interest and can cook that one easily.

The problem of interpretation is complicated, too, by the vocabulary problem. Consider the item "actress," which appears in one of the common inventories. The student is required to respond to this item by indicating like for, dislike for, or indifference to that word alone. The word covers a wide variety of performers from burlesque queen to moving picture star, to television bit performer, and to others. The advantages and disadvantages of being an actress of each of these kinds must be known to few of the responders, the pay range must be a question of speculation, and the opportunities in the profession must be known to a limited few. Without such knowledge the subject is to express like for, dislike for, or indifference to something about which she knows little. Yet that response is supposed to be used to help her to make one of the most important decisions of her life, the selection of a career. The argument that such items do differentiate occupational groups in general and may therefore be useful regardless of the subject's understanding of the items will be treated later. The suggestion that the objection to the inventories on the basis of vocabulary can be taken care of by providing a brief vocabulary list is ludicrous in its simplicity because a vocabulary list inclusive enough to incorporate all the definitions, their subtle shadings, and their variability would require weeks of study. As has been suggested, just one of the words, "actress," could be studied in all its ramifications and details for many, many hours. It is discouraging to find that some authors who have pointed out the vocabulary difficulties

involved have suggested only that the wording be changed rather than that the instruments be discarded.²³

Even the vocabulary problem would not be so serious were it not for the fact that there is great unevenness in the knowledge of the various items to which a subject must respond or choose among. In one preference record, high school students, even junior high school pupils, must choose the best liked among the following items: (1) sort and catalogue a valuable stamp collection, (2) write a popular article on how a diesel engine works, (3) determine the cost of manufacturing a new soap.

The first may have been a subject of lifelong concern to a counselee and he may know a great deal about it. To the others he may not have given a moment's consideration before taking the inventory and he knows little about them. Since it is likely that he will have some little knowledge about most of the items his responses will not, of course, be entirely random. He may have enough knowledge to produce a high or low score but not enough to produce a meaningful one. He is forced to make a choice among the items and this may require selection between ignorance and knowledge. *The item in which possession of information has meant reduction of glamour may be rejected in preference to one in which there are retained ignorance and continuing glamour.*

Even if one assumes that there is almost equal knowledge of items there may be vast differences in amount of concern about them. A counselee may want to indicate that an item deals with one of the most important events in his experience but cannot express his enthusiasm for it more forcibly than he can indicate a lukewarm interest. Johnson comments on this in a short article that deserves more attention than it has received. He says: "During my tenure of punch counting I observed that some marks were decisive and vigorous; others were hesitant and barely penetrated the paper. Some were bull's-eyes and others appeared on the periphery. Why

²³ Edward C. Roeber, "A Comparison of Seven Interest Inventories With Respect to Word Usage," *Journal of Educational Research*, September, 1948, 42:8-17.

a configuration of faint punches near the circle's edge should count as much as those accurately made and with sufficient force to mar the table and blunt the pin is indeed hard for me to understand. It seems rather apparent that a count of perforations is a pretty crude measure of interest."²⁴

The counselee may want to answer, when asked if he likes something, "Yes! Yes! Yes!" and to another question a hesitant, "Y-e-s," but he must circle the L's on the inventory in exactly the same way as if the differences did not exist. Research seems to show that extended opportunities to show enthusiasms add little to scores and do complicate scoring, but enough research on the importance of strong single interests has not been done simply because enthusiasms have been smothered in the multiplicity of items to which the person must respond. In counseling, the importance of real enthusiasm cannot be ignored or given minor consideration. Devices that do not provide for their expression cannot be effective in any phase of the counseling process.

It is commonly said that, though the total scores on inventories may be questionable, the separate items may be of some value in giving leads for interviews with counselees. It should be obvious, if the points given above are clear, that the statements about the questionable validity of the total test score apply equally well to the individual items and that single items may give as false leads as do the total scores. Further, however, it should be pointed out that the shotgun approach of throwing hundreds of items at counselees must often miss. To cover all the items that may be of particular significance to each counselee would require the use of thousands of items, and the significance of the one or even a few that might be important to him could easily be lost in the multitude. In what interest inventory can the boy who devotes all his time to the flora and fauna of his surroundings really express his interest? How can the boy who is set upon being a butcher or baker

²⁴ G. Johnson. "Mediations on an Interest Inventory." *Occupations*, February, 1952, 30:357-358.

express his enthusiasm? In what inventory will the girl who wants more than anything else in the world to be a trainer of horses indicate her interest? These are samples, and not unusual ones, of the kinds of counselees who cannot express their real preferences and who by being required to answer all the items must seem to indicate interests that are foreign to them.

Counselors who need such crutches for their interviews as responses called from long lists of inventory items ought to consider seriously whether or not they are skilled enough to stay in their profession. When a counselor has to turn to the general, the remote, and impersonal method of the interest questionnaire to get his interviews going and keep them moving he is likely to miss the *specific* and *close* and *personal* things that are really important to this particular counselee. He is likely to miss that essential factor in good interviewing, the encouragement of the counselee to reveal himself more fully than he can possibly do by circling symbols or punching holes.

In view of the pleasure that students get in taking the inventories, one might keep them for entertainment purposes—except for the fact that it seems impossible to prevent students from misinterpreting the results. Try as hard as a counselor will to prevent it, many students will believe that they have taken aptitude tests, employment tests, or vocational fitness tests. Having heard about scientific tests of aptitude and, in their ignorance, having been impressed with their value, they insist on reading into their interest inventory scores something that was never intended. The very nature of the inventories, their scores and norms, encourages this kind of misinterpretation. It is most unfortunate that inventory authors and publishers have not taken enough action to prevent the misuse and misunderstanding that are common among those who take the questionnaires.

On every page, perhaps even temporarily after every item if one must use the inventories despite all that has been presented on the preceding pages, it should be indicated that the items do not meas-

ure aptitude or ability. The counselor should teach his counselees to keep repeating the directions for taking the inventory when they answer each item and those directions will usually tell them that they should "assume that they have the training and experience (that may mean as much as five to ten or more years) necessary for all the activities." If that alone does not inhibit them from making too broad interpretations, maybe the further direction that they are to choose, "*What would you do as a regular thing if you were equally familiar with all the activities?*" will do so. Even the dull-est counselee will realize that he cannot be *equally* familiar with all the items.

It is probably true that many persons who use interest inventory scores have very little insight (though the authors of the tests may have) into the questionable assumptions and statistical procedures that are utilized in their construction, scoring, norming, and attempts to provide evidence of validity. Many of the procedures assume that the characteristics that are alleged to be measured are normally distributed despite the very obvious evidence that social pressures continuously skew such distributions and local mores, economic factors, nepotism, and endless other factors distort them. Do the users of the inventories realize that the items the subjects do not mark contribute, in a way they never intended, to certain areas because the forced choice residuals are scored? Do they really agree that the individual is atomistic in nature, that the atoms are independent and that they can be lumped together by simple addition to produce a really meaningful whole? Have they forgotten the lessons they learned in elementary school about the need of having something in common among the added items to make the sum meaningful? Do they know what kind of curious but involved reasoning is implied when it is suggested that scores above the seventy-fifth percentile are significant? Are they aware that there may be great overlap between two groups of individuals whose means differ significantly in, say, clerical and scientific interest? Do they know that much of the so-called validity data can be pro-

duced by the responses of that relatively small proportion of students in a total population who have such obvious interests that no measurement is necessary?

It does not seem likely that interest inventories will be easily routed from the scene because students do seem to enjoy filling them out. One teacher of English recently indicated that she had no faith in them but that she would not give them up for anything. "The two periods in which the pupils take the preference record, score it, and plot their profiles are two periods of the course that they like best." There is something fascinating in the popping noise that results when they push a pin through the stiff paper and it's fun, students say, to dream about being actors and astronomers and psychologists. Many of them have not had so much fun with pins and paper since they left kindergarten and it is a welcome change from school routine. Maybe they should not be denied the entertainment, but counselors ought to recognize that, until more convincing data about their value are available, interest inventories do not offer much more than that. Then they will have to consider whether the fun is worth the price.

VALIDITY OF PERSONALITY APPRAISAL TECHNIQUES

In the sections immediately above the discussion has centered around the interest inventories but much of what has been said is equally applicable to the so-called tests of personality. Examination of the literature about personality questionnaires and controlled interview techniques such as the Minnesota Multiphasic Inventory and the Rorschach reveals the same kinds of superficial reasoning noted above. A jump is made from platitudes about the importance of being adjusted, of being a well-balanced personality and being free from neurotic tendencies, to the suggestion that these things can be assessed in a dependable manner by responding to a long list of questions in lists or cards, or to ambiguous stimuli

in the form of pictures or ink-blots. Samples of such statements follow.

Insistence on respect for the "wholeness" of the adjusting organism, or guidance of the whole student, represents a major contribution of the modern movement in education. This personality test is an implement or tool through which the teacher can more easily and effectively approach this desirable goal. (*Manual of Directions for the California Test of Personality—Secondary Series*. Los Angeles: California Test Bureau, 1942, p. 1.)

The busy professional counselor and the guidance-minded classroom teacher have indisputable need for assistance in evaluating the various facets in the personalities of their counselee's ability level and interest pattern alone will not explain all the ramifications of individual behavior. . . . Fully to appraise the complete individual, therefore, one must of necessity include the pertinent personality factors in the equation. The Personal Adjustment Inventory has been designed to assist in this process. (*Manual for the Heston Personal Adjustment Inventory*. Yonkers, N.Y.: World Book Co., 1949, p. 1.)

Each of us has relatively permanent personality characteristics or traits known as our temperament. These aspects of personality are important for an understanding of ways we will act in school or industrial situations . . . we need a schedule that emphasizes important, stable traits which describe how normal, well-adjusted people differ from each other. The Thurstone Temperamental Schedule was devised for this purpose. . . . Seven areas of temperament are appraised in a relatively short questionnaire. (*Examiner's Manual for the Thurstone Temperament Schedule*. Chicago: Science Research Associates, 1950, p. 1.)

Following such generalized statements in the manuals, there is usually a series of dreary statements about validity that would strain the credulity of anyone who has had even minimum acquaintance with that topic. One is given coefficients of correlation with tests that are said to have been *previously validated* (*Manual for the Bernreuter Personality Inventory*); evidence that the inven-

tory does *not* do what it purports to do with attempts at explanations (Manual for the Thurstone Temperament Schedule, p. 10); low coefficients of correlation with self-ratings and those of other persons (Manual for the Heston Personal Adjustment Inventory, pp. 27, 28), and various other timeworn procedures that pronounce loudly to all who will listen that validity of these techniques is questionable.

Perhaps the continued widespread use of these devices is due in large part to the desire of many persons to get some data quickly. It has been pointed out by several writers that such devices vainly seek the pot of gold at the end of the rainbow—a simple, cheap, foolproof method for studying human personality. But fools continue to rush in where angels fear to tread and there seems to be no immediate hope that the waste of time and money on such instruments will soon be reduced. Though research on human personality should not be discouraged, it seems clear that nothing revealed by the study of behavior suggests that there is promise in the continued use of the standardized questionnaire or inventory method.²⁵ Whether or not the counselor agrees with these latter statements, he will be forced to agree that there are no data on the predictive validity of such instruments that justify their current use in counseling.

VALIDITY OF PROJECTIVE METHODS

Discouraged by the lack of evidence on the validity of structured personality measurement devices, the counselor may be tempted to turn to projective methods (*not* tests) in the hope that he will get some valid measurement of personality. After he has examined the evidence, his discouragement may turn to complete disillusion-

²⁵ R. B. Cattell states the situation well in these words. "The only situation in which a questionnaire would have complete validity would be a person with complete integrity and complete self-knowledge—and such a person would scarcely need a personality test." Raymond B. Cattell, *Description and Measurement of Personality*. Yonkers, N.Y.: World Book Co., 1946, p. 343.

ment.²⁶ Attempts to validate such devices in all the usual ways has resulted invariably in negative results or in findings of such low relationships between projective scores and criteria that their use for diagnosis and prediction is profitless. (It is possible that psychiatrists and clinical psychologists can use such methods as aids to diagnosis in seriously disturbed cases, but counselors in educational institutions are not usually clinical psychologists or psychiatrists.) Projective methods have been used in the study of such groups as obese women, blind adults, stutterers, adoptive parents, discordant marriage partners, children with reading disabilities, unsuccessful students, applicants for admission to various kinds of training, and many other groups. In most of such studies, which seem more like advertisements than experiments, it is difficult to separate what has been merely claimed from that which has been amply demonstrated. In all cases the results offer the counselor little to help him in his duties.

Projective methods raise all the usual difficulties in validation of personality measurements mentioned at other places in this volume. The use of confusing language in reports of subjects' performances on the projective instruments makes their interpretation particularly difficult.

The following report on a student was submitted after he had taken the Rorschach. It has been edited only in the sense that the actual symbols of categories used by the clinician have been removed.

RORSCHACH REPORT ON A COLLEGE STUDENT

This is the Rorschach picture of a personality that is almost completely constricted emotionally. He represses all tendency to respond emotionally either to the stimuli of the world outside

²⁶ J. W. M. Rothney and R. A. Heimann, "Development and Applications of Projective Tests of Personality." *Review of Educational Research*, February, 1956, 26:56-71. See also related chapters in the *Annual Reviews of Psychology*. Stanford University Press.

him or to promptings from within (fantasy, creative thought). Probably severe emotional shocks in early conditioning left him afraid to respond emotionally to others, to trust them or love them, for fear they might hurt him or let him down. Yet his dependence on such external values is complete for he never trusted himself enough to develop what inner resources he had into values of his own creation.

His intelligence does not seem to be more than average (perhaps lower than college average) and this lack of superior intellectual ability makes any form of adjustment difficult for a constricted person unless he is living in a group with low standards of intelligence and achievement. (Frequently the highly constricted person has superior intellectual capacities which enable him to attain accepted cultural values in the area of adjustment in society.) But for this young man, three value-areas are closed off—that of inner values of his own making, that of outer values of emotional relationships with others, and that of outer values of intellectual achievement. For him life is an insoluble conflict between his dependence for security on the achievement of conventional outer values and his inability to attain them because of excessive distrust of himself and of others (*inferiority feeling*). He is very suspicious of others, particularly of people in positions of authority. He is afraid of being imposed upon, afraid something will be put over on him, afraid above all that he will make a mistake, that he will be unable to understand and meet expectations, and that his status and security will be further diminished by his failure. Any unfamiliar situation where he is not sure of the outcome and has to take a chance is terrifying to him. He is especially afraid of any competitive or examination situation. He does not dare to depart from immediate perceptions of reality, to generalize on the basis of past experience, and to adapt to new situations on that basis. This clinging to immediate reality is illustrated by his repeated questions during the Rorschach test: "Are these supposed to represent animals or anything?" "Were these supposed to represent something?" He cannot get over the idea that the blots represent something in

reality which he is supposed to identify. Any notion of selecting aspects of the blots for himself and outlining forms resembling those he has seen in past experience is difficult for him to hold.

He is very confused about himself and why he cannot succeed in his endeavors as others do, but he has failed to so often that he has no confidence, always expects the worst, and because of his fears is unable to utilize what capacities and information he has. Compulsively driven to try to meet conventional standards, e.g., of academic achievement he defends himself at the same time, against his fear of failure, by a childish sort of halfway opposition; i.e., frustrated, he directs his aggression outward. He hasn't the courage or self-confidence to carry his oppositional impulses very far into action but is sporadically resistant and submissive—will contradict but smile as he does. This outward expression of hostility is one of the means by which he manages to keep functioning in the face of his severe neurotic conflict. It is probable that if he directed his hostility toward himself any more than he does, he would become completely depressed, losing what stamina he has to continue his battle.

A second means by which he helps make life bearable consists in a thought and action pattern of escape. He dodges and looks sideways at every problem as a whole for fear it might be beyond him. Instead, in his confusion, he focuses attention on some small part which he thinks he can handle safely, and thus, frequently overlooks the obvious and important, in favor of the obscure and inconsequential. His academic work probably reflects this habit of thought, as well as a tendency to make hasty, inaccurate judgments in moments of defiance when he wishes to get clear of a trying situation. His anxiety blocks his memory and makes verbalization difficult.

He is, however, intelligent and rational enough to know, most of the time, when he is not attaining standards expected of him in both academic and social life, but he does not know why and feels confused and helpless. His occasional awareness that many of the satisfactions others find in life do not exist for him, leaves

him with a sense of emptiness and futility. Although such feelings have a depressing effect, they are healthy insofar as they indicate that he is not reconciled to his abnormal way of life.

Perhaps the description given above may not be representative of what projectionists commonly report, but if there is a difference it simply emphasizes the fact that variation in reporting methods makes the problem of getting dependable data more acute.

The counselor may wonder whether the above was worth the expenditure of approximately four hours for administration, scoring, and writing the report. It is said that the student is "almost completely constricted" but no definition of "almost" is given. He is said to repress "*all* tendency to respond emotionally," which seems to be an impossible situation. It is said that "probably severe emotional shocks in early conditioning left him afraid to respond emotionally to others" but that statement permits the additional one that *probably* it did not. It is said that "his intelligence does not seem to be more than average" and one would wonder if there are not better ways of estimating intelligence (undefined here) than by use of the Rorschach method. With the above suggestions for critical analysis of the report the reader may wish to continue his appraisal of it and try to decide whether it would be useful in counseling.

The writers are in sympathy with the general projective idea and believe that it may be used informally and with locally devised materials,²⁷ but they find that scores derived from the standard instruments defy interpretation. The difficulties are described in the following passage.

Projectionists do not have any common metric comparable to the difficulty level concept used by achievement and intelligence testers on which their items can be scaled. Most of their "tests" must be administered individually and the time element becomes important.

²⁷ J. W. M. Rothney and Bert A. Roens, *Counseling the Individual Student*. New York: Dryden Press, 1949, pp. 132-134.

It may require about four hours to administer, score, analyze, interpret, and report a Rorschach. When that time is contrasted with the unlimited number of test administrations that might be given with a group test, it will be seen that projective norms are harder to establish, and one can understand why the number of subjects in projective research is frequently small. Statisticians have not yet developed satisfactory techniques for treatment of the variables and relationships which projectionists profess to abstract from their data. And it is still impossible in the very nature of the projective situation to untangle the test administrator from the score.²⁴

It appears, then, that validity of projective techniques has not yet been developed to a point at which they can contribute significantly to the work of the counselor even though they *may* be of value for the work of the clinical psychologist or psychiatrist.

RELIABILITY OF INTEREST AND PERSONALITY MEASUREMENTS

Although the sections on validity in the manuals of interest and personality inventories are usually short and devoid of data, one commonly finds long sections on reliability. These may be impressive to many test users who are still confused about reliability and validity and who tend to interpret reliability coefficients as indicating dependability. The instruments might be more useful if the time, energy, and space that were devoted to computing and reporting coefficients of reliability were spent in further study of validity.

The usual methods of computing reliability for achievement and aptitude tests have been carried over directly to the fields of interest and personality measurement, but there is some reason to doubt that the conditions are similar enough to warrant such carry-over. The split-half method described in Chapter III requires that the items in each half be of equal variability and of *the same quality* as those in the other half. It is extremely doubtful that the items

²⁴ J. W. M. Rothney and R. A. Heimann, "Development and Application of Projective Technics," *Review of Educational Research*, February, 1936, 26:56-71, p. 66.

on inventories can be of the same quality to a particular individual. A subject may be consistent in his circling of the L (for like) after such items as actor, astronomer, fat men, and snakes, but it seems rather dangerous to assume that these have the same quality to him for all of the reasons given above in the section on validity of interest inventories.

Since very few interest or personality inventories have two or more parallel forms, the method of administering two forms and computing the correlation between scores on them is seldom used. The coefficients obtained from test-retest method with varying intervals between two administrations of the inventories generally lie between .50 and .70. They are not high enough to permit accurate forecasting of an individual's score at a later time.²⁹ When one finds studies using this method it would seem well to remember, in interpreting them, that they report consistency of scores, not necessarily stability of interest or personality.

The test-retest method produces what some persons have called a *coefficient of stability*. It has long been recognized that memory of items and responses and general "test-wiseness" in the test-retest method produces higher coefficients than are obtained by other methods. In personality and interest inventories it seems likely that the memory factor would be more influential than in achievement testing because many of the items may be more emotionally charged. There is always, too, the problem of variability in moods from time to time.

Use of the test-retest method in the study of reliability of interest tests may be as useless as trying to determine the reliability of a thermometer by checking the readings made at one hour against those made at a later hour during a day. The answers to questions

²⁹ See such studies as the following: R. Jacobs, "Stability of Interests at the High School Level." *Educational Records Bulletin*, No. 52, 1949, pp. 83-87. W. K. Trinkans, "The Permanence of Vocational Interest of College Freshmen." *Educational and Psychological Measurement*, Winter, 1954, 14:641-646. W. L. Lyton, "The Variability of Individuals Scores Upon Successive Testings on the Minnesota Multiphasic Personality Inventory." *Educational and Psychological Measurement*, Winter, 1954, 14:634-640.

on an inventory about feeling of belongingness on a day when family circumstances have been ideal for a youth may change significantly if he happens to be questioned the day after a family quarrel has resulted from his insistence upon borrowing the family car.

All the above considerations and those previously reported in Chapter III must raise some doubt in a counselor's mind about the meaning and value of the so-called coefficients of reliability that he finds in manuals for interest inventories and questionnaires. It seems unlikely in view of the lack of validity data that he would ever use these instruments, but, if he should be tempted to do so, the absence of clarity about stability of scores should inhibit him completely.

NORMS OF PERSONALITY AND INTEREST INVENTORIES

In the pages above, the data on validity and reliability of inventories and questionnaires have been examined and have been found inadequate. When further questions are raised about their norms, chaos is added to confusion.

Some authors of inventories and questionnaires have carefully avoided the presentation of norms but others present elaborate norm tables. Still others, while they do not use the term, imply that there are norms by suggesting that one can get an A, B, or C by marking an inventory in a manner similar to the way that members of an inadequately defined group mark them.

During the early stages of the guidance movement, the process of vocational counseling was described as one of putting the square peg in the square hole. It was suggested that certain occupations required certain kinds of personalities and capacities and that the job of the counselor was to find the persons with suitable patterns of both, and to steer them into suitable occupations. This kind of thinking is retained by some authors and users of inventories who seem bent on obtaining measurements of the individual's interest

patterns or personality traits and directing the individual (subtly, of course) into an occupation or into training for it. To such persons an array of scores obtained from an inventory, a questionnaire, or a controlled interview is necessary. They want a score that is quickly obtained and readily transmitted into a letter grade, percentile rank, or other converted score so that it can be used for educational vocational guidance or personality therapy.

It should be clear to the counselor that there are no satisfactory measures of vocational interests (the name, "Strong Vocational Interest Blank," for example, is simply a christening by its author), preferences, or personality characteristics of any occupational or educational group. And it should follow very clearly that, until such measures are available, the norms provided with personality and interest questionnaires are useless for counseling purposes. The elaborate profiles that are produced from the scores are socially, psychologically, and mathematically unsatisfactory. Tyler has expressed this very well in the following paragraph, although it is unfortunate that she uses the word "test" when she means inventory or questionnaire.

It seems to be commonly believed by vocational counselors that personality tests can be used to ascertain whether an individual has the special traits required by a certain occupation. Is this boy dominant enough to be a good salesman? Is this girl well adjusted enough to be a social worker? Unfortunately there is almost no research evidence that warrants our using personality tests in this way. There is, on the other hand, a considerable body of evidence showing that such use is unwarranted. There are two things lacking. In the first place *we do not know what the personality traits essential to the various occupations are.* In the second place we cannot be sure that the inventories measure what we think they do. This means that when a counselor tells a client, "This test shows you are a dominant person who should be able to succeed as a salesman," *it is much worse than telling him nothing at all and leaving him to make a decision on other grounds.*

The disadvantages of bringing them (personality measurements) in may well outweigh any possible advantages. The essence of this kind of counseling is perfect candor; the counselor does not withhold information from the client. But what is one to tell a person who makes T-scores of 70 or higher on the Hs, D, and Pt scales of the Minnesota Multiphasic Inventory? Even if it were good therapeutic practice to state that he showed an unusually high level of the personality characteristics found in hypochondriacs, depressives, and obsessive-compulsive neurotics—and it obviously is not what any therapist would approve of—there is *not even sufficient scientific evidence to warrant such a statement.* A better procedure would be to phrase the meaning of the scores in simple, nontechnical language and say, "This test shows that you are more likely than the average person to be concerned about symptoms of possible illness, to feel discouraged and depressed at times, and to get ideas and impulses that you can't shake off," and then use this statement as a starting point for further discussion. The value of the test results to him, however, is still very doubtful. If the interviews encourage free expression the symptoms that are complicating his life will come up for consideration eventually anyway. All that the test has accomplished by bringing them out in the beginning in this form is to add to the load of anxiety he is carrying.³⁰ (*Italics added.*)

As indicated in the above quotation, it would be difficult to interpret norms even if it were possible to have some that were based on sufficient numbers of well-described subjects. The difficulty is compounded when the subjects are insufficient in number and inadequately described in terms of such factors as age, sex, education, religion, social class, rural or urban residence, geographic location, intelligence test scores, occupation, marital status, health, and the scores of other factors that may determine responses on interest and personality assessment devices. The tendency to use adults confined to mental institutions as basic groups, against which to compare controlled interview or questionnaire responses of young persons

³⁰ Leona E. Tyler, *The Work of the Counselor*, New York: Appleton-Century-Crofts, 1933, pp. 134-135.

who are not, has encouraged some weird interpretations of scores and some strange distortions of terms that have been previously attached to severe maladjustments. The complex problem of developing suitable norms in personality and interest measurement seems not likely to be solved by the use of captive audiences (one author has described a large segment of them as an "atypical minority of studious, docile, and intelligent humanity which sits in university classrooms") as basic groups whose scores are to be used in comparisons with those who are free.

In a report on norms in general the following statement appears:

Unfortunately, many alleged norms reported in test manuals are not backed by even an honest effort to secure representative samples of people in general. Even tens or hundreds of thousands of cases can fall woefully short of defining people-in-general. Inspection of test manuals will show (or would show if information about the norms were given completely) that many such massed norms are merely collections of all the scores that opportunity has permitted the author or publisher to gather easily. Lumping together all the samples secured more by chance than by plan makes for impressively large numbers; but while seeming to simplify interpretation, the norms may dim or actually distort the counseling, employment, or diagnostic significance of a score.³¹

If this statement is true about the relatively simple measurement of achievement and so-called aptitude, and it seems to be, the counselor should consider how much more dimming and distortion the inadequate norms for the personality questionnaires and interest inventories produce.

SUGGESTIONS FOR THE COUNSELOR

In view of the many limitations of attempts to measure interest and personality, it would seem well for the counselor to eschew them completely. Indeed, it would seem desirable to have authors

³¹ Harold G. Seashore and James H. Ricks, Jr. "Norms Must Be Relevant." *Test Service Bulletin* No. 39. New York: The Psychological Corporation, 1950, pp. 16-17.

and publishers agree to declare a moratorium on their production for, say, a 20-year period or until the time that researchers or philosophers or both could, by concerted efforts, develop more satisfactory instruments. (Just how some counselors would spend their time and how journals would keep up their publications without some scores to manipulate and report would present new problems.) Measurement in any field must always follow a long period of description and it appears that we are only in the descriptive phases of the study of personality and interests.

While counselors are waiting for better instruments it would seem desirable to heed these words of one author: ". . . Those who have real professional training will not need a system. Those who lack psychological knowledge will help pupils more effectively by using simple human warmth and interest than by thumbing a handbook of oversimplified recipes."²²

Counselors should consider, too, the words of still another writer presented below.

It is imperative that we come to understand some of the forces that are driving children to act the way they do. It seems to be an absolute necessity for us to try to think that learning is not only positive in the sense that one can facilitate it by understanding the laws of learning, but learning can often be facilitated greatly by removing things that can be thought of as blocks to the process.

Somehow or other, through this kind of working with children (informal methods) they do feel a release; sometimes the greatest form of progress is made, not by banging away at the intellectual task, but by taking one's time at the emotional task and coming somehow or other into rapport with these children. Under these circumstances they come to feel they are respected, wanted, and loved, and that they are missed if they are absent. They come to have self-respect and, hence, develop respect for others. They release many talents which have been hidden. They grow!²³

²² L. F. Shaffer in O. K. Buras, *The Third Mental Measurements Yearbook*. New Brunswick, N.J.: Rutgers University Press, 1949, p. 70.

²³ L. E. Rath. "Understanding the Individual Through Anecdotal Records, Sociometric Devices and the Like." *Goals of American Education*. American Council on

The statement, and the article from which it is drawn, are not just sentimental statements but *hard-headed recognition of the fact that assessment of interests and treatment of personalities are complex processes that cannot yet be done quickly, impressionistically, or mathematically.*

If a counselor has been using the questionnaires or inventories and has had some misgivings about their use, he may want to try some other methods of getting meaningful information about his counselees. There *are* other ways of getting such information and most of them still seem more promising than the standardized instruments in this area.³⁴ Among them are systematic observation and reports by means of behavior descriptions obtained from those who have had sufficient opportunity to observe counselees in a variety of situations; cumulative reports of participation in organizations; reports of selection of activities, courses, units, or topics when choice is permitted; selection of jobs when several are available; and interviews specifically designed to discover interests that are meaningful rather than transient or whimsical. Finally, of course, if a counselor feels that he must have some sort of lists to which his counselees are required to respond, he can make up his own in a form that permits local references and adaptations,³⁵ that does not force choices where the counselee has no real choice, permits expression of enthusiasms, and encourages him to state that he has insufficient information on which to respond. In any case he should be sure to provide the opportunity for his counselees to tell *in their own words* and in the length and detail of their own choosing what their interests, feelings, and problems really are.

SUMMARY

In this chapter it has been suggested that there are no valid

Education Reports No. 40. Washington, D.C.: American Council on Education, April, 1950, pp. 63-73.

³⁴ John W. M. Rothney and Bert A. Roens. *Counseling the Individual Student*. New York: Dryden Press, 1949, Chapter III.

³⁵ A sample of a locally devised inventory is presented in Appendix I.

short cuts to the appraisal of personality, attitudes, interests, and behavior of counselees. Examination of the form and content of self-descriptive inventories, records, blanks, and projective devices has revealed so many shortcomings that their use by counselors cannot be recommended. (No judgment has been made about their use by clinical psychologists or psychiatrists.) Claims for the value of such instruments are either unsupported or the evidence that is offered is inadequate. It has been suggested that the study of various aspects of the complex behavior of individuals must undergo a long period of descriptive study before valid measurement in this area can be established. Until this is done it has been suggested that the counselor should employ more direct and personal methods of studying the behavior of his counselees by use of observation, behavior description, interview, and analysis of performance techniques.

EXERCISES

1. It is generally agreed that self-report inventories are easily faked. What implications does this have for the actions of the counselor who is seeking data in the area of interests or personality?
2. The Science Research Associates have recently presented a new form of the Kuder Preference Record, Form D. Compare and contrast this new form with the previous forms on such factors as:
 - a. Empirical evidence of validity.
 - b. Reliability information.
 - c. Norm groups presented for comparisons.
3. Take any five commonly used interest inventories and in a five-column chart list their stated purposes. In a second column find and list the author's stated theory of interests.
4. As a counselor you have been asked by your school board for data on the number of "maladjusted" students in your school. What procedures would you employ in seeking the answer to this question?
5. At a conference of counselors you hear the statement, ". . . I

know [name of an interest inventory] has poor validity, but we use it in our school because we find that the students like to take it and we feel it motivates them to think about occupational choice. . . ." Discuss the implications of such a statement.

6. Select five personality inventories and list the personality traits or characteristics presumably measured by each. How are these defined? Are the definitions consistent from inventory to inventory? What factors might account for the inconsistencies? What inferences would you draw from this exercise regarding the measurement of personality characteristics?
7. Comment critically on the following statement.

The psychologist stands aghast at the self-assurance with which professional school counselors in America diagnose the personality faults of little children and the boldness with which they undertake the delicate task of adjustment. . . . The student of genius who is familiar with the motivating influences that have their origins in quirks of childhood personality shudders to think what the results would have been if school counselors had had a chance to "adjust" the personalities of the budding geniuses of history. One can imagine them, freed from all their peculiarities and complexes, adjusted to the world as it was and becoming indistinguishable from the common herd.

8. The following letter was actually written to the parents of a high school student by a counselor. Comment critically on it.

I am enclosing a report on the Strong Vocational Interest Test for Women which was recently given to your daughter. As you can see, her highest scores are in the business area, where she received two "A" ratings. Also to be considered, I think, is a "B minus" rating as buyer, which might fit in very well with the business interest. She is showing some teaching interest at the elementary level, but definitely not at the high school level. You can see that she also has a good rating as a housewife! Even so, her interests so far as vocation is concerned are less feminine than the average girl. By feminine interests I mean interest in the esthetic, cultural, and in the more personalized things. For instance, she would

- ment" in L. L. Thurstone (Ed.). *Educational Applications of Psychology: Essays to Honor Walter V. Bingham*. New York: Harper, 1952.
- Hamlin, Roy M. "The Clinician as Judge: Implications of a Series of Studies." *Journal of Consulting Psychology*, August, 1954, 18:233-238.
- Kuder, G. Fredrick. "Expected Development in Interest and Personality Inventories." *Educational and Psychological Measurement*, Summer, 1954, 14:265-271.
- Lindzey, Gardner. "Thematic Apperception Test: Interpretative Assumptions and Related Empirical Evidence." *Psychological Bulletin*, January, 1952, 49:1-25.
- Rothney, John W. M., and Heimann, Robert A. "Development and Applications of Projective Techniques." *Review of Educational Research*, February, 1956, 26:56-71.
- Rothney, John W. M., and Schmidt, Louis G. "Some Limitations of Interest Inventories." *Personnel and Guidance Journal*, December, 1954, 33:199-204.
- Stordahl, Kalmer E. "Permanence of Strong Vocational Interest Blank Scores." *Journal of Applied Psychology*, December, 1954, 38:423-427.
- Strong, Edward K. Jr. "Permanence of Interest Scores Over 22 Years." *Journal of Applied Psychology*, April, 1951, 35:89-91.
- Windle, Charles. "Test-Retest Effect on Personality Questionnaires." *Educational and Psychological Measurement*, Winter, 1954, 14:617-633.

CHAPTER IX

The Future

The treatment up to this point may be regarded indirectly as an overall description of the current state of testing for counseling in the general secondary school program. The questions asked, the doubts raised, the criticisms implied, and the limitations described should make it clear that much remains to be done if the testing program is to be improved.

If one were to take 1916, the year in which the Stanford-Binet was published, as a base year, and appraise the accomplishments in testing for counseling since that time, the results would be very discouraging. The measurement in guidance movement that began with such high promise has failed to meet the expectations that had been raised by its auspicious start. When counselors seek tests that meet the standards that measurement experts have themselves called essential, the quest is generally unrewarding. Currently they can get only approximations of what they need. Although they may hope to get better instruments in the future, it does not appear that they will get them soon.

One author who has had much to do with testing has described the current test situation very effectively in the following words:

What about these tests? Are they all worth using? Consider this unorthodox system of classification: (1) There are obsolete, once useful tests which would die gracefully but for the inertia of test-users who still want them. (2) There are some ill-conceived tests which should never have been born and some which have matured badly; these remain alive because of the psychometric innocence of some test-users. In deference to the proud authors I shall name none. (3) Some excellent tests are useful for very limited and special purposes and will never have widespread usage. (4) Certain modern, well-standardized tests have and should continue to have wide-spread application. Revisions and improvements can be expected. (5) Some newer tests are still frankly experimental and can be used cautiously by those who understand their background and present status.¹

Even the items numbered (3) and the first part of (4) may be overly optimistic, but the frankness of the whole statement by one who is concerned with the construction and distribution of tests is commendable.

Another author has been equally pointed in his appraisal of the current situation in testing. While writing essentially about personality measurement, his comment is applicable to measurement in other areas and appears particularly so for the field of counseling. "Up to now most of us have been more ambitious to exhibit the gold extracted from the deep earth of personality by this or that cathected technique than we have been to discover, explain, and, if possible, remedy the failures and limitations of the method. The truth is that during this first phase of our common enterprise we have not known enough about our subjects to estimate the validity of every technical rating or statement that was made. And, furthermore, we have not been disposed to criticize unreasonably."²

From the purely historical and developmental standpoint, it may

¹ Harold G. Seashore. "Understanding the Individual Through Measurement." In Arthur E. Traxler (Ed.), *Goals of American Education*. American Council on Education Studies. No. 40. Washington, D.C.: American Council on Education, 1950.

² Foreword by Henry A. Murray in Harold H. Anderson and Gladys L. Anderson. *An Introduction to Projective Techniques*. New York: Prentice-Hall, Inc., 1951.

appear to some that the authors of this volume have been *disposed* to criticize unreasonably. It may be said, for instance, that compared with other sciences, centuries old, psychological testing is but an infant. There is strong temptation to "go easy" with infants but we cannot go easy on testers at the expense of the youth to be counseled. From the standpoint of *time*, it can be agreed that psychological testing is indeed an infant, but the data regarding numbers of tests currently used strongly suggest that the infant is extremely large for its age and growing rapidly.

What is really of great concern at this point is that, like other infants, psychological testing *will* grow up. The answers to the questions "How will it grow up?" and "What will it become?" must wait the passage of time, but the discussion that follows suggests some of the possible developments.

SOME POSSIBLE DEVELOPMENTS

No one can foretell accurately what the future of testing and test development is likely to be. The possibilities are varied; some are remote, others seem less so. No one of the possibilities discussed below is likely to operate in isolation and whether anyone will ultimately dominate remains to be seen.

Major developments in testing seem likely, at this time, to take place in areas other than in their application to counseling. This, to a considerable degree, has been the case in the past and it seems likely that it will continue to be so in the future. The motivation for better development will probably come from forces outside the school counselor's office—from the armed forces with their problems of selection and from the manpower shortage of industry. The results of development of testing in these areas, as suggested in the opening chapter, have been felt in the schools and counseling offices of our nation and it seems likely that they will continue to be felt.

But will it just be more of the same? It seems likely that we can

look forward, at least in the immediate future, to what might be described as "more of the same" with unsatisfactory new tests appearing periodically as they have for the past quarter of a century. As with many new tools, these products are likely to be in the same pattern with similar built-in weaknesses and limitations as those that have characterized their predecessors. This trend may be tempered to some extent by the serious efforts of *some* authors and *some* publishers to meet the more rigid standards that appear to be emerging. Some factors that may influence trends are discussed below.

MORATORIUM AND LEGISLATION

One possible way to improve tests would be to declare a moratorium on the production of new testing instruments for a period of years. During the period obsolete and poorly conceived tests would be killed off, gains would be consolidated, and standards of test design and marketing might be strengthened. Unfortunately, such a plan can hardly be taken seriously, for the attractiveness of profits from test production and marketing is too great for many persons to resist. It does seem strange, however, in a country in which butchers' and grocers' scales are regularly checked and policed, and clothiers' tags of "100 percent wool" must be validated if the sellers of such products are to avoid imprisonment, that a test distributor may sell his products without any supervision or regulation. After reading many test manuals one is often left with the feeling that "there ought to be a law." And it may come to that. It seems to have been assumed in the past that educational or psychological tests could be produced and distributed without any kind of regulation. It has been found necessary to enforce compliance with Pure Food and Drug Acts to protect even professional persons, who, presumably, should not need protection. Perhaps educators, psychologists, and counselors need similar legislation

for protection from those who have taken advantage of freedom from control.

THE INFLUENCE OF RAPID CALCULATORS

Current trends toward automation, development of rapid computers, and the applications of punched card procedures seem likely to influence the direction in which test construction and utilization may proceed. While conditions may change rapidly, the current international situation is such that mobilization of both military and civilian skills is a matter of great concern and methods of effective utilization of manpower are constantly being studied. Wesman, discussing the topic of what is new in guidance testing, reported an investigation that could not have been conceived as possible without elaborate computation procedures in the following words: "Two government activities perhaps deserve first mention because of their very scope. The first is a project being directed by R. L. Thorndike for the Air Force, the ultimate purpose of which is to conduct an aptitude census of the American people. Eventually, this program would presumably result in cataloguing the population with respect to its abilities, just as the well-known censuses have catalogued it with respect to age, income, material possessions, education level, etc." ²

The manpower shortage in many areas and in many skills has been viewed with alarm in many quarters and the efforts of many are being expended to seek out talent and to inventory potential. Educators who have attempted to define their role in this manpower shortage frequently advocate the use of large-scale testing programs in the schools. In a recent publication of the Educational Policies Commission, one of the implications for education of the manpower situation is directed to guidance in the following recommendation: "*Improved guidance and counseling.* Guidance serv-

² Alexander G. Wesman, "Guidance Testing," *Occupations*, October, 1951, 30:10-14.

ices, uniquely characteristic of American Education, should be further improved, and so increased in scope as to involve all who teach and to reach all who learn. *Guidance programs should be soundly rooted in understanding of the manpower situation.*"⁴

The rising tide of school enrollments, unprecedented in our history, may well be counted among the forces which, in a time of increasing automation, can precipitate a mass testing movement that could make our present efforts appear meager.

Each of these forces has a "mass" dimension about it and solutions are likely to be sought through techniques in which great masses of data are gathered. The possibility of losing individuals in a mass of punched cards does not seem remote. This thought was envisioned by Hull, at least in part, as long ago as 1923. Because of its current appropriateness, his description of "a machine which makes aptitude forecasts automatically" is quoted in full.

Another system of making aptitude predictions from forecasting formulae has been devised by the writer in the form of an automatic machine. This machine is an integral part of a comprehensive program of vocational guidance first sketched in 1923. The program calls for the construction of a single universal battery of tests which shall sample, so far as possible, all of the important aptitude determiners. The battery will contain perhaps thirty or forty different test units and require a day or more to administer. Upon the basis of this one battery there will be constructed separate forecasting formulae for each of the more important type occupations—possibly to the number of forty or fifty. Thus there would be forty or fifty different equations, each equation weighting the tests of the one battery in a different way so as to make the best possible forecast of a particular aptitude. These equations would, of course, all be much longer than the one given above for freehand drawing, each probably involving every one of the thirty or more tests of the battery. To make all the forty or fifty forecasts by such a system in the ordinary way would

⁴ Educational Policies Commission, *Manpower and Education*. Washington, D.C.: National Education Association, 1956, p. 126.

involve something like 1,500 multiplications, all of which would need to be summated in a more or less complicated manner. This would represent a huge amount of labor, to say nothing of the human errors certain to creep into such a large amount of hand work. The forecasting machine mentioned above has been designed to perform this work automatically. Three of these machines have been constructed.

In its final form, this machine will have the different forecasting formulae placed in it permanently as a four-inch perforated band of thin metal, somewhat resembling a music roll in appearance. The test scores will be given the machine in the form of a similar perforated band of paper upon which have been recorded the test scores of a given subject. The test scores are recorded on this paper band by means of a special perforating device which is operated something like a typewriter. A series of forty test scores may be thus recorded in about a minute. Once the test scores have been recorded on the paper band, it will be placed in the forecasting machine and the starter pressed. The machine will then proceed automatically, and without any attention whatever from the attendant, to make one aptitude forecast after another until the entire forty or fifty have been calculated.

At the time of inserting the band of test scores there will also be placed in another part of the machine a card bearing the names of the subject and a blank form giving in a column the names of all the aptitudes and occupations for which forecasting formulae are available. As the machine makes its forecasts, it will stamp them down on this card automatically, opposite the names of the appropriate aptitudes. When the forecasts have all been made the machine will stop automatically, at the same time ringing a bell to call the attendant. The card of forecasts, when removed from the machine, will then present in orderly array and in units of single uniform scale, permitting of instant comparisons, forecasts of the individual's probable success in all of the chief occupations of the world. The youth whose potential aptitudes are thus recorded may then examine the card to learn those vocations in which his chance of success is low. These may be avoided in his choice of life work. He may then

examine the card to learn those vocations in which his chance of success is greatest. The three or four most promising vocations thus emerging may be given further investigation. From these, in the light of his interests, opportunities, and general circumstances, may finally be chosen a life work.

It scarcely needs to be pointed out that the program of vocational guidance thus briefly sketched is a revolutionary departure from the current development of aptitude testing. This being the case, there will no doubt be considerable inertia and resistance from conservative quarters. To this difficulty must be added the fact that the program involves a vast amount of minutely coordinated research quite impossible of accomplishment by isolated workers. But the logic of the situation is certain to triumph in the end. We may look forward with confidence to a day not far distant when some such system as that sketched above will be operating in every large school system. Then, and not until then, will there be possible a genuine vocational guidance for the masses of the people.⁵

Much of what Hull envisioned in 1923 has materialized. One of the components for his "comprehensive program of vocational guidance" is available in several batteries of tests. Dvorak's description of the assumptions underlying The General Aptitude Test Battery published by the U.S. Employment Service in 1947 matches very closely the idea that Hull expressed. She says: "The basic assumption underlying the GATB is that a large variety of tests can be boiled down to several factors and that a large variety of occupations can also be clustered into groups according to similarities in the abilities required. This makes it feasible to test all of a person's vocational abilities in one sitting and to interpret his scores in terms of a wide range of occupations."⁶

A battery of tests devised by Flanagan is designed for much the same purpose. His battery is described as follows in a publication of the Science Research Associates. "*The Flanagan Aptitude Classi-*

⁵ Clark L. Hull. *Aptitude Testing*. Yonkers, N.Y.: World Book Co., 1928, pp. 487-490.

⁶ Beatrice J. Dvorak. "The General Aptitude Test Battery." *The Personnel and Guidance Journal*, November, 1956, 35:145-154.

fication Tests (FACT) comprise the core of the aptitude batteries in the Job-Test chart. These tests have been designed to provide measures of aptitude for 14 critical job elements. Two qualities of the FACT series appear to commend its use in programs of personnel selection; (1) The test items are highly similar to tasks performed by workers in business and industry. (2) Flanagan's job-analysis studies have indicated recommended combinations of tests corresponding to the requirements of particular occupations." (Science Research Associates. S.R.A. *Job-Test Chart*, 1954.)

The computations involved in such approaches are no longer a problem when digital computers are capable of handling almost astronomical figures. The IBM type 650 computer, known as a "magnetic data processing machine," could handle Hull's 1,500 multiplications with consummate ease. It can make 5,000 multiplications or 3,700 divisions in a minute, 78,000 additions or subtractions of ten-digit numbers in the same length of time.

It will take but a little imagination on the part of the reader to envision the establishment of computer centers to which counselors could forward data obtained by test batteries to be punched on cards and fed into a machine such as the one described above. A machine programmed to operate on the basis of critical scores could then feed results to appropriate forms that would be returned to the counselor.

It takes but little more imagination to visualize an extension of the above wherein copies of all these results could again be placed on punched cards and deposited in centers where they could be sorted mechanically for the purpose of assigning manpower in cases of emergency. The obvious limitation in all this mechanical processing is the fact that the machines can utilize only the data that are fed into them—they cannot improve on them. And the data, at this time, must be scores from imperfect tests.

The danger inherent in the prospect described above is, of course, that counselors may find themselves increasingly victims of expediency—of large numbers and of urgency. Instead of working

more and more in the direction of individualization, counselors may succumb to group concepts underlying such approaches as described above and may find themselves utilizing techniques designed for purposes other than work with individuals. One of the major tasks of the counselor as a consumer of tests, then, will be that of resisting the temptation to adopt for his own use methods and techniques that were developed for other purposes.

CONSUMER DEMAND AND TEST IMPROVEMENT

Tests and use of tests may be improved as the result of increasing sophistication and consequent demands for higher standards by users. The discouraging lack of progress to date may have occurred because test users have not learned what to seek from test publishers. Their failure to require high-quality tests has resulted in huge production and sales of instruments that cannot possibly do what is claimed for them. It does not seem likely that there will be a significant improvement until the level of sophistication of test users has been raised. As Seashore has pointed out, "sponsoring the growth of good tests is a joint operation of all of us who are test-consumers and test-producers." He indicates some signs of improvement in these words: "Considerable progress along these lines is occurring. The professional societies and technical groups are beginning to write more rigorous standards for test construction, standardization, and validation. And what is more important, the level of sophistication in measurement is rising among educators, guidance personnel, and psychologists. Increasing consumer discrimination is a potent club over test authors and their publishers."¹

The discussion above of four factors that may influence the future of developments in testing is not inclusive. With the exception of the suggested moratorium, mentioned more, perhaps, as an expression of wishful thinking than serious speculation, each of

¹ Seashore, *op. cit.*

the developments may occur singly or in subtle combinations. Ultimately it would seem to the authors that the best prospect for the future will lie in the increased sophistication of test users. Such development seems likely to be slow.

SOME BASIC PROBLEMS IN TEST DEVELOPMENT

Demand for higher standards by test users should, in the long run, result in the production of better tests than are now available. Before they can become really useful, however, there are some basic problems that must be solved. The most difficult one is that of developing satisfactory methods of quantifying human behavior. In building a test it is common practice to select enough items to fill up less than one hour of testing time. Each of these items is then, by some mysterious process, allotted an equal value on a scale. These scores on a scale are often totaled by a "scorezee" or "quick-scoring" device and the total is christened as "mental maturity," "mechanical reasoning," "mental ability," or such words as currently suit the contemporary pedagogical jargon. As such it is proclaimed as a better measure of an individual than the "subjective" judgment of a teacher who has observed a pupil daily in his classes for a year or more. At times it may be, but the claim that it is usually true lacks convincing evidence. Before such evidence can be obtained it seems essential that the following problems must be solved.

NEED FOR EXAMINATION OF BASIC CONCEPTS

It is conceivable that a test with high predictive validity might be constructed by selecting items subjectively, scoring them as though each item were of equal value, and christening the total score. Currently, however, no one can really claim that an instrument with high predictive validity has been produced by such methods. The best results obtained so far are indicated by such

small coefficients of correlation between test scores and criteria that prediction of an individual's later performance in the area christened by the author is little better than chance.

Perhaps the time has come for test builders to reexamine their basic premises and techniques. Continuation of the usual timeworn processes of test construction that have proved to be almost sterile seems not to be justified. There appears to be little hope of significant contributions to counseling by workers in the testing movement until a testing Einstein arrives to shake up its very foundations. The current way to appear scientifically and statistically respectable is to follow the beaten path and to grind out again, with perhaps minor refinements, what has been endlessly ground out before. Sorokin, in an article that should be read by anyone who considers the use of tests, states the case very effectively in these words:

Obsessed by metromania, our testers indefatigably measure their test data and present them in an "exact" and "objective" form of numerical scores, indexes, statistical tables, marvelously decorated with impressive looking mathematical formulae and other simulacra of a precise quantitative research. Manufacturing of these "quantitative movies" is done so artfully that many a logically and mathematically innocent onlooker seriously takes this sham-quantitative appearance for a genuine reality. A legion of psychosocial researchers sincerely believe that these impressive looking scores, indexes, rows of figures, coefficients of correlation, probable errors, standard deviations, coefficients of reliability, and so on, deliver; but the objectively studied and exactly measured "diamonds" are but arbitrary, subjective, often fantastic, assumptions of the testers dressed up in quantitative costumes and mechanical make-ups. Our testing numerologists have as far relationship to real mathematics as had various numerologists and astrologers ("mathematici" as they were called) of ancient times and of the middle ages.^a

^a Pitirim A. Sorokin, "Testomania," *Harvard Educational Review*, Fall, 1955, 25:199-213.

THE PROBLEM OF DEFINITION

The problem of securing agreement among test producers and consumers about definition of terms has limited, and seems likely to continue to limit, the contributions of testing to counseling. The definition of intelligence has always been subject to dispute and the differences in definition have not been resolved by substitution of the words "scholastic aptitude." The word "aptitude" seems to have almost as many definitions as there are authors of aptitude tests if one may judge by the kinds of items that appear in their tests. Rulon has pointed out some of the problems and some of the implications for counseling that result from the failure to clarify definitions used in testing in the following words:

There appears to be relatively complete confusion as to whether ability and aptitude are things which are changing or things which do not change. Certain aptitudes are goals to be attained, while others are determiners of goals. We try to make choices on the basis of some abilities, and we try to develop certain other abilities such as to get along with people and to use the scientific method. We even speak of developing the *ability* to read. We believe that if a person cannot be a good stenographer easily, we should encourage him to be something else. But we believe that if a person cannot be a good citizen easily, then we should accept the responsibility of developing him into a good citizen *anyway*. Clearly the concept of achievement in relation to ability needs clarification, and we should decide more definitely what should be done about individual aptitudes. We need a ready answer to the inquirer who may ask, "Why don't you improve the aptitudes you find this individual deficient in? In the case of legal aptitude why not develop it, just as we try to develop social aptitudes?" *

Rulon goes on to point out that lack of clarity in definition of

* Phillip J. Rulon, "On the Concepts of Growth and Ability," *Harvard Educational Review*, Winter, 1947, 17:1-9.

terms makes a very great difference in actual counseling procedures. He writes:

Here is a case stated in two ways by two different educational workers. One says, "We have many cases of discrepancies between ability and ambition in one way or another: either more ambition than ability or more ability than ambition." The other worker says, "Here is a boy who has failed all his academic work throughout the grammar grades, and is now failing in all his academic subjects in the junior high school, and yet he says his plans are to go into a bookish occupation after college training."

These two workers agree that they are presenting the same educational problem in different words. Let us see what the ambiguous *entity concept* of ability will lead us to in this case. We call the boy in and say, "Johnny, your ambitions and your ability do not jibe." The boy must not be expected to be very happy about this, nor very enlightened.

Now let us see to what the consistent performance definition of ability leads us. We say to the boy, "Johnny, you know what your record is and what you have said your plans are. Now I want you to consider these things in relation to each other, and I want you to do something about them. You can do one of three things as I see it. You can change your level of performance, you can change your plans, or you can expect to have lots of trouble. Your performance is not consistent with your plans."

If school personnel will heed warnings such as those given by Rulon, there will be less confusion about the value of tests in the counseling process. There will be less criticism of students because they are "not working up to their ability" and more understanding of that curious phenomenon, the boy who is "working beyond his ability." There will be less tendency to attach labels to students and assume that they have been permanently classified. There will be more realization that test scores do not, in themselves, provide sufficient data for counseling and more appreciation of the fact that

they do not automatically determine the action that a counselor or his client should take.

SUBJECTIVITY AND OBJECTIVITY

Perhaps little progress in the testing movement can be made until test users stop leaning too heavily on the word "objective." Rothney and Roens have commented on this problem in the following words:

The previously noted tendency to make all educational procedures seem standardized, scientific, and statistical has obscured the fact that most objective techniques have required a good deal of subjectivity in the process of their construction. Examination of test manuals reveals that all the authors have made some subjective judgments. If the authors of the manuals took their items from textbooks, they had to *choose* among many whose authors had, in turn, made *subjective decisions* about the materials they selected. The authors of tests must have made *judgments* concerning the kind and number of items to be selected, and even if they used statistical criteria to guide them, they had to *decide* which of many criteria they would use. They had to *choose* their scoring methods and attach values to the items. Again, if they were chosen by statistical procedure the authors had to *make selections* from many. They made *choices* among criteria against which to validate their tests and they were required to *decide* how far they would go before they were convinced that validation was adequate. The "objectivity" of a test is found in scoring procedures (again derived largely by *subjective techniques or choice* from so-called objective techniques) that make it possible for two scorers to get the same results. The student who is aware of all these problems will not be too easily influenced in the selection of techniques by the fact that one technique is said to be objective and the other subjective. Completely objective techniques (except in the very narrow concept of scoring items) for obtaining a minimum list of items to use in counseling are not now available.¹⁰

¹⁰ John W. M. Rothney and Bert A. Roens, *Counseling the Individual Student*. New York: Dryden Press, 1949, pp. 85-86.

It seems unlikely that there will be much progress in the testing movement until test users begin to recognize the oversimplification, confusion, and misrepresentation that occurs when an "objective" label suggests to them that tests have greater exactness and scientific rigor than they do.

THE EMPHASIS ON SPEED

As one reads test manuals and literature about tests one must observe that speed seems to be of the essence. The California Test Bureau in a cover design that came "from the offices of Raymond Loewy Associates, internationally famous designers of such products as the Studebaker car, Schick razor, and Frigidaire refrigerator," announces in screaming type that its "scorezee" test is EASIER of administration, BETTER for interpretation and FASTER in scoring. Otis' tests seem attractive to many because they are QUICK-SCORING. Many authors in their test manuals emphasize the fact that the test can be administered during one class period. The reason for so much emphasis on getting the testing over with as rapidly as possible is difficult to ascertain. If the performance of pupils is hurried, if the sampling of items has been limited so that they can be done in one school period, and if scoring is to be done so quickly that thorough checking and examination of pupils' answers is not possible, one must wonder how dependable the results can be.

When test users require test builders to put *ease* and *speed* of administration or scoring second to validity, reliability, or adequacy of norms, testing for counseling may become a more useful procedure. The provision of machine-scored tests, although it has relieved much drudgery, has resulted in overemphasis on speedy "objective" testing. An observer might conclude from examination of our tests that the major objective of American education was to develop true-false and multiple-choice minds and to do it *fast*.

THE STATISTICAL SITUATION

It is difficult to determine whether discussion of this topic should be placed in the section on limiting or promising factors in the improvement of the testing situation. It seems that at some time in the near future educational statisticians must discover that they are working with human beings rather than plants in fertilized plots, or balls drawn out of jars. Until they begin to realize that they have the special problem of dealing with the most complex thing in the world—a human being growing in very complex circumstances—there seems to be little hope of practical contribution of statisticians to the counseling process.

There can be no doubt that statisticians have elaborated the obvious fact, probably known by instructors since teaching began, that there are individual differences among groups of students. Whether the esoteric language used in the elaborate quantification of the facts of individual differences has contributed significantly to improvement of counseling of students is a question on which there could be much difference of opinion. When the counselor attempts to use the products of the statistician he finds that, by the time he has allowed for the assumptions involved, made adjustments for differences in circumstances of his own counselees, discounted the difficulties in conversion of raw scores to theoretical scales, and tried to translate the resulting product into language that a counselee or his parents can understand, he has little that has been worth the effort.

And statisticians, by and large, seem to be more concerned with fleeting moments in a child's life than with his development over a period of time. Of the thousands of research studies only a very few have used longitudinal data. The usual practice is to take a quick sample of a child's behavior with a test that takes less than an hour of his time, and then to hurry the scores to the computing machines for correlation or for one of the currently popular modes

of analysis—factor, cluster, discriminant, or dispersion. In the process the data are normalized or distorted in some way so that they can be manipulated more easily. The findings usually indicate the general characteristics of a group of subjects and it is often implied that they are rather permanent. The results are then published and the researcher turns away to another problem with another group of subjects whose test scores are manipulated in a new way that is currently popular. Statisticians rarely undertake a thorough study of individuals over a long enough period of time to determine whether or not their findings have significance for those who work with growing human beings. Not until their studies encompass many data about persons over longer periods of their growth does it seem likely that they will contribute significantly to the processing of data for counseling purposes.

Changes in practices in a profession *may* mean that there is growth. Sometimes it means only a floundering process that results in retrogression. At still other times it may mean only faddism that distorts and destroys. The historian who has looked at the measurement scene must have noticed the popularity of the correlation coefficient, chi-square, analysis of variance and covariance, factor analysis, and discriminant analysis in turn. As one becomes popular it is applied to every possible kind of data, regardless of their origin and the purposes for which they were first derived. Borrowing always from agriculture and other rather simple sciences to get under their halo, the educational statisticians apply their methods to the complex study of human beings and bring little of value to the counselor. The difficulty of translating experimental designs dealing with soil fertility, weight of pigs, or effectiveness of manurial treatment into a guidance framework has not yet been resolved.

TESTING AND FREE SCHOOLS

The literature of testing suggests that testers would find their

best working conditions in a regimented rather than in a democratic state. In a country in which a state or federal administration regulated the curriculum, the hours to be devoted to it, the policies for pupil placement, promotion, or retardation, and in a nation in which workers could be allocated to jobs without respect to their personal wishes, the testing movement might be highly effective. As "national" norms are now presented it is assumed that every subject's scores can be compared in a reasonable manner to those of a supposedly representative group despite the exceptional assets or liabilities in the situation in which he finds himself. Some forty years after the testing movement began one organization is finally planning to study the influence of local curricula, school offering, and quality of teaching on test scores.¹¹ It should be noted that the Educational Record Bureau has used separate scores for private schools for many years.

In the United States local schools are given much freedom in choosing the educational opportunities they will provide, and it is known that great differences in offerings result. The process of comparing a youth's scores with a set of national norms, and making deductions about him on the basis of the comparison, implies that he has had comparable learning opportunities. To make such comparisons when it is obvious that he has not had such opportunities seems hardly to be a sound procedure. In a regimented state where markedly similar opportunities could be ordered, and where individuality was not important, the testing movement would seem to serve well. In view of the great differences in practices in American education, however, the professed claims of test authors that testing aids in better placement and guidance of students seem to lack merit. Until the test authors provide test users with data that will enable them to give proper weight to such factors as the quality of instruction and the kind of home, school, and community experiences to which the testee has

¹¹ *Educational Test Service Developments*. Vol. III, No. 4, Princeton, N.J.: Educational Test Service, May, 1955.

been exposed, there seems little hope that tests can really be used effectively.

It is probable that, when the testing movement comes of age and testers recognize that differences in educational opportunities will always appear in a free society, test manuals will contain a statement such as this: "The following factors contribute to the achievement of certain scores on this test. They contribute to those scores in these proportions." This statement will be followed with a weighting of such variables as quality of instruction and instructional materials, amount of time spent on the subjects, interest or motivation, and the physical condition of the student. Until such data are available, however, the counselor must rely on his judgment about the significance of a test score. Escape from this dilemma does not seem imminent.

CURRENT SIGNS OF PROGRESS

It has been indicated several times throughout this volume that prediction of behavior of humans is a hazardous procedure. Accordingly no attempts will be made to predict what testers will do. One may, however, note some of the current activities that seem to indicate promise of improvement. Four of them are listed and discussed below.

1. Publication of new tests that indicate more careful attempts at standardization and validation.
2. Publication of standards for the sale and distribution of tests.
3. Provision of methods by which critical reviews of tests may reach the potential consumer.
4. Raising of standards for training and employment of guidance workers.

PUBLICATION OF BETTER TESTS

At times, despite all the difficulties mentioned above, some rays

of hope appear. One of them is the trend away from omnibus tests of intelligence, mental maturity, or mental ability and toward assessment of several kinds of performances by means of a test battery standardized on the same population.¹² It appears that the era of the "Intelligence Quotient" is now rapidly passing.

The lumping of performances into one whole score called an "IQ" has served the very useful purpose of making educators aware of individual differences in children for so long a period that it is now ready to be pensioned off. Scores on several separate performances are likely to take the place of the IQ on cumulative records. This is not a new development since several writers have for a long time advocated the breaking down of omnibus scores. In the haste in which testers always seem to be, however, they have taken such small samples of student performance that the scores are not likely to be dependable. In the future it seems that there will be more use of batteries of tests with titles that clearly indicate their content, and will sample enough of a student's various performances sufficiently to permit comparisons.

Comparison of the manual of the Differential Aptitude Tests¹³ published in 1947 and one of the older single score tests, such as the Henmon-Nelson Tests of Mental Ability,¹⁴ published in 1932, reveals some startling contrasts. The data on the older test were presented in two pages but 77 pages are required to present several hundred validity coefficients, standard error and reliability data, and fairly adequate description of the norms and standardization populations of the newer test. The authors of the Differential Aptitude Tests have also published a follow-up over a seven-year

¹² The use of the term "Intelligence Quotient" without the word "score" attached to it has been most unfortunate. The addition of the word might have reminded those who used the term that all they ever obtained from a test was a fallible score, not an infallible index of true mental organization.

¹³ George K. Bennett, Harold G. Seashore, and Alexander G. Wesman, *A Manual for the Differential Aptitude Tests*, New York: The Psychological Corporation, 1952.

¹⁴ V. A. C. Henmon and M. J. Nelson, *Teacher's Manual for the Henmon-Nelson Tests of Mental Ability*, Yonkers, N.Y.: World Book Co., 1932.

period that purports to establish the long-term predictive efficiency of its subtests.¹⁵

The pattern of complete reporting of the D.A.T. has been followed with a greater or lesser success by several succeeding "multiple aptitude tests," some of which, like the Multiple Aptitude Tests of Segal and Raskin,¹⁶ attempt to provide "differential intelligence score" norms in addition to the usual grade and sex norms. Both of these tests have encouraged the user to employ the concept of expectancy tables as a way of interpreting the predictive validity of their scores.

Other innovations among recently published tests include such items as attempts to establish occupational ability patterns in The General Aptitude Test Battery of the United States Employment Services;¹⁷ the use of a standard score reporting system in the Multiple Aptitude Tests; the plan to continue the revision of norms and continue validation studies with the same group of boys and girls used in the preliminary validation studies presented in the Flanagan Aptitude Classification Tests;¹⁸ the purported comparable growth scores in the School and College Ability Tests; the profile reporting scheme of the School and College Ability Tests that allows for graphic presentation of standard error concepts;¹⁹ the thorough norm data of the revised Stanford Achievement Tests;²⁰ the effort to measure essay writing and listening comprehension in the new Sequential Tests of Educational Progress;²¹ and

¹⁵ George K. Bennett, "The D.A.T.—A Seven-Year Follow-Up," *Test Service Bulletin*, No. 49. New York: The Psychological Corporation, November, 1955.

¹⁶ David Segal and Evelyn Raskin. *Manual for the Multiple Aptitude Tests*. Los Angeles: California Test Bureau, 1955.

¹⁷ Beatrice J. Dvorak. *General Aptitude Test Battery*, Washington, D.C.: Department of Labor, United States Employment Service, 1947.

¹⁸ John C. Flanagan. *Aptitude Classification Tests, Technical Supplement*. Chicago: Science Research Associates, 1953.

¹⁹ Coöperative Test Division. *Examiner's Manual for the Coöperative School and College Ability Tests*. Princeton, N.J.: Educational Testing Service, 1955.

²⁰ Truman L. Kelley and Others. *Stanford Achievement Tests*. Yonkers, N.Y.: World Book Co., 1953.

²¹ Coöperative Test Division. *A Prospectus for the Coöperative Sequential Tests of Educational Progress*. Princeton, N.J.: Educational Testing Service, 1957.

reports to their users that go beyond a sales promotion approach and furnish information concerning general measurement concepts as well as empirical studies dealing with particular test scores and their uses. Typical documents available upon request to the publisher include the Test Service Bulletins and the Test Service Notebook series of the World Book Company, and the Test Service Bulletin series of the Psychological Corporation.³⁹

Local school systems are attempting to improve the use of testing materials by trying out tests in their own schools and by providing in-service training programs in measurements and guidance. Englehart has quoted a letter that shows how this may be done. It is reproduced below.

In each elementary school I [Miss Eloise Cason, Child Guidance Director, Bloomfield, N.J., Public Schools] have a meeting with the staff and/or new teachers on how to "read" a permanent record card. Records of the children in the school are duplicated and used as a basis for discussion. It is my impression that the average teacher learns most effectively about the use of tests by focusing on their meaning for understanding a *particular* child. Test data are related to other information. It is also possible at these meetings to do a bit of educating on validity, reliability, etc. Similar meetings are held at other school levels.

In the junior high school the guidance staff prepares summaries of test data to help the teacher plan for the *group* with which she is working.

When new tests are introduced in skills or the content fields, the teachers concerned are asked to evaluate the suitability of the material in the tests in relation to their program. I believe this procedure is greatly appreciated, and removes some of the fear of tests.⁴⁰

³⁹ See, for example: Roger T. Lennon. "A Glossary of 100 Measurement Terms," *Test Service Notebook*, No. 13. Yonkers, N.Y.: World Book Co. Marie P. Dolansky. "Predicting Success in College," *Test Service Bulletin*, No. 80. Yonkers, N.Y.: World Book Co.; Jerome E. Doppelt. "How Accurate Is a Test Score?" *Test Service Bulletin*, No. 50. New York: The Psychological Corporation, 1936.

⁴⁰ Max D. Englehart. "Testing and Use of Test Results," *Review of Educational Research*, February, 1956, 26:3-13.

RAISING STANDARDS OF GUIDANCE WORKERS

The coming of age of a professional group is usually indicated by the introduction of processes of accreditation or certification, by the development of professional literature and research at an increasingly high level of competence, by recognition of established related organizations, and by increased acceptance on the part of the public. All these processes seem to be in operation in the counseling movement. As they develop it seems likely that the workers in it will become intelligently self-critical about their activities. In the process it seems likely that the selection and use of tests will receive much consideration.

The trend toward increased professionalization and more adequate preparation of school counselors seems clear when it is noted that 34 states now offer counselor certificates and at least seven other states are in the process of drawing up such regulations.⁴¹ Analysis of the specific requirements for these various state counseling certificates points up the fact that over 90 percent of them include an area of "Analysis of the Individual" among their requirements. They usually require at least one course dealing with the use of tests and measurements in counseling the individual. As more and more counselors-to-be have these preprofessional training experiences it is hoped that their sophistication in the use of test results will rise.

Other signs of increasing professional awareness among guidance workers include the publication in 1953 of the report entitled *Ethical Standards of Psychologists* by the American Psychological Association and the first draft of a parallel code of ethics by the American Personnel and Guidance Association in 1957. These guides may help the counselor to raise his professional status and

⁴¹ See: Royce E. Brewster, "Guidance Workers' Certification Requirements," *Guide Lines*, Washington, D.C.: U.S. Office of Education, 1956, 43 pp. Arthur J. Jones and Leonard M. Miller, "The National Picture of Pupil Personnel Services in 1953," *The Bulletin of the National Association of Secondary-School Principals*, February, 1954, 38:103-159.

to remind him of the limitations of his technical tools in working with counselees.⁴²

Both the American Personnel and Guidance Association and the American Psychological Association have influenced training requirements for various counseling fields. The latter association has recommended that a two years' master's degree be offered to train "psychological technicians" at the subdoctoral level, and has stressed the need for supervised practice in the use of the various psychological tools including tests. This pattern might well influence training of guidance workers, for there is much dissatisfaction with the present levels of preparation in this field.⁴³ At a workshop held in conjunction with the national convention of the American Personnel and Guidance Association and attended by counselor-trainers from all parts of the country, several areas in the training of counselors were considered.⁴⁴ It was agreed that more actual experimental data were needed before a definitive program of training could be outlined. A similar conclusion was reached by Stoughton. In writing on the preparation of counselors he said: "The scarcity of research on inservice training and growth should be viewed as a harsh challenge to a profession in which knowledge has increased so rapidly and in which there are known to be many workers whose professional training is, to say the least, limited."⁴⁵

It appears that in the next few years much research and experimentation in the areas of determining counselor competence and

⁴² American Psychological Association, *Ethical Standards of Psychologists*, Washington, D.C.: American Psychological Association, 1953, 171 pp. American Personnel and Guidance Association, *First Draft of a Proposed Code of Ethics for the American Personnel and Guidance Association*, Washington, D.C.: American Personnel and Guidance Association, 1957.

⁴³ American Psychological Association, "The Training of Technical Workers in Psychology at Subdoctoral Levels," *American Psychologist* September, 1955, 10:541-545.

⁴⁴ National Association of Guidance Supervisors and Counselor Trainers, *Workshop Report*, Washington, D.C.: American Personnel and Guidance Association, April, 1957.

⁴⁵ Robert W. Stoughton, "The Preparation of Counselors and Personnel Workers," *Review of Educational Research*, April, 1957, 27:174-185.

the kinds of preprofessional training experiences that will help him arrive at that stage of preparation will be done.

THE SKEPTICAL COUNSELOR

Having looked to the future and critically evaluated the present in regard to the use of tests in counseling, just where do counselors stand? Do they "throw the baby out with the bath water" and disregard all test data until that day somewhere in the hazy future when all reliability coefficients will be 1.00 and validity evidence will be reported in coefficients of at least .97? Or do they admit that these tools are not perfect and plan to continue to use them because they are the best available at the present time? Perhaps there is another way to avoid these extremes of complete rejection or unsophisticated acceptance.

The way may be found in the development of a healthy skepticism, a critical and challenging search for data that offer evidence of increasing efficiency of the tools with which they work. It is a skepticism that would look critically at nontest as well as test approaches to the understanding of the individual counselee. It would not reject a wise and tempered "subjectivism" in favor of blind and empirical "objectivity" simply because the latter uses the language of mathematics and the former the seemingly less precise language of description.

It may be the skepticism of a counselor who has learned to question the blatant, merchandising claims of some test publishers and who insists that the authors of tests furnish proof that they do at least approximately what they claim to do. It is a skepticism that encourages the counselor to make his own studies, to work out his own validity, reliability, and normative data based upon test performances of students in local school systems and to follow his counselees throughout their school careers and on into post high school training and work. It is a skepticism that takes nothing for

granted, and keeps him seeking dependable devices to help his counselees to help themselves to make sound decisions about their plans and actions.

SUMMARY

In this chapter an examination of the current status of the testing movement and some projection into the future have been attempted. The examination has shown that test authors are much concerned with achieving statistical sophistication, but seem less concerned with basic problems of rationale and validation. It has been suggested that the pace of production of new tests based upon timeworn procedures should cease until a careful examination of the very basis of present-day psychometrics has been made. Future developments in test construction and design have been considered and some promising current "new" procedures have been discussed. The greatest hope for the immediate future was seen in increased consumer sophistication as antidotes to high pressure sales promotion campaigns of test publishers. This sophistication and increased professionalization among school guidance workers was seen as necessary if counselors are to find tests useful in the important tasks that they undertake.

EXERCISES

1. Examine the catalogues of five major test publishers. Compare their stated policies on restriction of sales of tests to potential buyers.
2. Compare the test manual published in the 1920's with one published in the 1950's. What are the major differences?
3. In a chart compare and contrast the strong and weak points of three modern "multiple aptitude tests." Use the criteria for test evaluation developed in Chapter V for this exercise.
4. Contrast the basic philosophical differences in point of view of

guidance programs in England and France with typical guidance programs in the United States.

REFERENCES

- American Educational Research Association. "Educational and Psychological Testing," *Review of Educational Research*, February, 1956, 26:1-110.
- American Educational Research Association and National Council on Measurements Used in Education. *Technical Recommendations for Achievement Tests*. Washington, D.C.: National Education Association, 1955. 36 pp.
- American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education, Joint Committee on Test Standards, "Technical Recommendations for Psychological Tests and Diagnostic Techniques," *Psychological Bulletin* (Supplement), March, 1954, 51:1-38.
- Anastasi, Anne. *Psychological Testing*. New York: Macmillan, 1954, 682 pp.
- Brewster, Royce E. "Guidance Workers' Certification Requirements," *Guide Lines*. Washington, D.C.: Office of Education, 1956, 44 pp.
- Buros, O. K. (Ed.). *Fourth Mental Measurements Yearbook*. Highland Park, N.J.: Gryphon, 1953, 1163 pp.
- Educational Policies Commission. *Manpower and Education*. Washington, D.C.: National Education Association, 1956, 126 pp.
- Ellis, Albert. "Recent Research with Personality Inventories," *Journal of Consulting Psychology*, February, 1953, 17:45-49.
- Kuder, G. Frederic. "Expected Developments in Interest and Personality Inventories," *Educational and Psychological Measurement*, Summer, 1954, 14:265-271.
- Ross, C. C., and Stanley, Julian C. *Measurement in Today's Schools*. Revised. New York: Prentice-Hall, 1954, 485 pp.
- Rulon, Phillip J. "On Concept of Growth and Ability," *Harvard Educational Review*, Winter, 1947, 17:1-9.
- Sorokin, Pitirim A. "Testomania," *Harvard Educational Review*, Fall, 1955, 25:199-213.

Thorndike, Robert, and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education*. New York: Wiley, 1955, 575 pp.

Travers, Robert M. W. *Educational Measurement*. New York: Macmillan, 1955, 420 pp.

APPENDIX

Activities Reports

It was suggested in Chapter VIII that requiring counselees to respond to long lists of interests, activities, or questions designed to measure personality characteristics is not likely to provide valuable data for counseling. The practice is, however, so common that it is not likely to be given up. In view of this situation it is suggested that counselors who feel that they must have such lists should make up their own in a form that permits of local reference, that does not force choices where the counselee has no real choice or information, permits him to state that he has had insufficient experience with an activity, discourages attempts to fake responses (since the items he marks are to be discussed during interviews), and provides opportunities to indicate activities that are very important to him.

The following Activities Report is presented to show how this may be done. All the listed activities were obtained from interview notes entered on the cumulative records of subjects of the Wisconsin Counseling Study.¹ The items were put into series so that they cover consecutively the areas of reading, sports, music, school clubs, out-of-school organizations, collecting, hobbies, care of animals, domestic duties, art, and miscellaneous. A general unclassified section appears

¹ John W. M. Rothney. *Guidance Practices and Results*. New York, Harper & Bros., 1958.

at the end of the lists, and a supplementary report on activities is required on the last page.

The counselor may choose to summarize a counselee's activities on the Analysis of Report on Activities in preparation for interviews.

The Activities Report is not intended for use as it appears here. It is offered as a suggestion of a method that counselors might use in their work. They may, of course, want to change the time intervals used in this sample. The effort to construct an instrument of this kind should reveal to the counselor some of the inadequacies of the expensive and unrealistic devices commonly employed. He *may* find that the instrument he develops is more useful in counseling than those he has purchased in the past.

ACTIVITIES REPORT

Name _____
City _____
Grade _____
Date _____
(month) (year)

The purpose of this exercise is to learn about what kinds of things you have done during the past year. Listed on the attached sheets are many of these things. Opposite each one are the four letters D, M, Y, N.

*D means almost every day

*M means about a few days a month

ACTIVITIES REPORT (Continued)

Y means only two or three times a year

N means you have not done this within the past
year

If you have done these things almost every day make
a circle around the D.

If you have done these things a few times a month,
make a circle around the M.

If you have done these things only two or three times
in the last year, make a circle around the Y.

If you have never done these things, make a circle
around the N.

*NOTE: Where the activity is seasonal, such as ice
hockey in the winter, answer in terms of
almost every day, a few days a month, two
or three times a year, or never during the
season.

ACTIVITIES REPORT (Continued)

EXAMPLES: Hanging around the local dairy bar (named) D M Y N
 Collecting stamps D M Y N
 Going to another state. D M Y N
 Working out coefficients of alienation D M Y N

You will have plenty of time, but work carefully. Mark every item. There are no right answers. Everyone will probably check the list differently. After you have finished, go back and underline any item that you consider very important to you. At the end of your lists there is space for you to write additional activities. Add as many things as you want. You might also want to describe some of the activities you have marked. At the next interview the counselor will talk to you about the items you have marked.

- 1R. Reading mysteries D M Y N
- 2S. Playing football. D M Y N
- 3M. Playing a musical instrument D M Y N

ACTIVITIES REPORT (Continued)

4K.	Attending meetings of school athletic organizations	D	M	Y	N
5K.	Attending meetings of a Mariners Club	D	M	Y	N
6C.	Collecting stamps	D	M	Y	N
7H.	Making model airplanes.	D	M	Y	N
8P.	Caring for animals.	D	M	Y	N
9D.	Sewing	D	M	Y	N
10A.	Painting pictures	D	M	Y	N
11X.	Riding motor bicycles	D	M	Y	N
12R.	Reading fiction	D	M	Y	N
13S.	Playing baseball.	D	M	Y	N
14M.	Playing in a band	D	M	Y	N
15K.	Attending meetings of school dramatic club	D	M	Y	N
16K.	Attending meetings of Scouts.	D	M	Y	N
17C.	Collecting matchbook covers	D	M	Y	N
18H.	Making model boats.	D	M	Y	N
19P.	Working with horses	D	M	Y	N
20D.	Doing housework	D	M	Y	N
21A.	Drawing pictures.	D	M	Y	N
22X.	Writing plays	D	M	Y	N
23R.	Reading sport stories	D	M	Y	N

ACTIVITIES REPORT (Continued)

24S.	Playing basketball.	D	M	Y	N
25M.	Playing in an orchestra	D	M	Y	N
26K.	Attending meetings of Y-clubs	D	M	Y	N
27K.	Attending meetings of DeMolay or CYO.	D	M	Y	N
28C.	Collecting coins.	D	M	Y	N
29H.	Making model railroads.	D	M	Y	N
30P.	Raising pigeons	D	M	Y	N
31D.	Cooking	D	M	Y	N
32A.	Drawing maps	D	M	Y	N
33X.	Writing stories	D	M	Y	N
34R.	Reading comic books	D	M	Y	N
35S.	Boxing	D	M	Y	N
36M.	Playing in small instrumental groups	D	M	Y	N
37K.	Attending meetings of Future Farmers of America.	D	M	Y	N
38K.	Attending meetings of Job's Daughters	D	M	Y	N
39C.	Collecting photos of movie stars.	D	M	Y	N
40H.	Building radio or TV equipment	D	M	Y	N
41P.	Raising rabbits	D	M	Y	N
42D.	Knitting	D	M	Y	N
43A.	Decorating own room	D	M	Y	N
44X.	Writing poetry.	D	M	Y	N

ACTIVITIES REPORT (Continued)

45R.	Reading adventure stories	D	M	Y	N
46S.	Playing tennis	D	M	Y	N
47M.	Attending meetings of a choir	D	M	Y	N
48K.	Attending meetings of 4-H Club	D	M	Y	N
49K.	Attending meetings of Sub-deb club.	D	M	Y	N
50C.	Collecting travel folders	D	M	Y	N
51H.	Building racing cars	D	M	Y	N
52P.	Raising chickens	D	M	Y	N
53D.	Embroidering	D	M	Y	N
54A.	Designing clothing.	D	M	Y	N
55X.	Operating motor bikes.	D	M	Y	N
56R.	Reading animal stories.	D	M	Y	N
57S.	Bowling	D	M	Y	N
58M.	Singing in groups.	D	M	Y	N
59K.	Working on the school paper	D	M	Y	N
60K.	Attending meetings of the club	D	M	Y	N
61C.	Collecting postal cards	D	M	Y	N
62H.	Making trout flies	D	M	Y	N
63P.	Showing cattle.	D	M	Y	N
64D.	Crocheting	D	M	Y	N
65A.	Arranging flowers	D	M	Y	N

ACTIVITIES REPORT (Continued)

66X.	Working with a chemistry set.	D	M	Y	N
67R.	Reading about ways people make their living	D	M	Y	N
68S.	Playing pool and billiards.	D	M	Y	N
69M.	Attending meetings of a church choir.	D	M	Y	N
70K.	Attending meetings of student committees	D	M	Y	N
71K.	Attending meetings of a yacht club.	D	M	Y	N
72C.	Collecting statues.	D	M	Y	N
73H.	Woodworking	D	M	Y	N
74D.	Weaving	D	M	Y	N
75A.	Painting	D	M	Y	N
76X.	Helping to build a house.	D	M	Y	N
77R.	Reading farm journals	D	M	Y	N
78S.	Playing pingpong.	D	M	Y	N
79M.	Attending meetings of a chorus.	D	M	Y	N
80K.	Attending meetings of a language club	D	M	Y	N
81K.	Attending meetings of a forestry club	D	M	Y	N
82C.	Collecting rocks.	D	M	Y	N
83H.	Building model cars	D	M	Y	N
84D.	Making clothes.	D	M	Y	N

ACTIVITIES REPORT (Continued)

85A.	Drawing cartoons.	D	M	Y	N
86X.	Working on guns	D	M	Y	N
87R.	Reading biographies	D	M	Y	N
88S.	Track	D	M	Y	N
89M.	Attending meetings of the glee club	D	M	Y	N
90K.	Attending meetings of a tumblers' club	D	M	Y	N
91K.	Attending meetings of the English club	D	M	Y	N
92C.	Collecting records.	D	M	Y	N
93H.	Building bird houses.	D	M	Y	N
94D.	Setting hair.	D	M	Y	N
95A.	Clay modeling	D	M	Y	N
96X.	Making things at home	D	M	Y	N
97R.	Reading love stories.	D	M	Y	N
98S.	Swimming	D	M	Y	N
99M.	Solo singing	D	M	Y	N
100K.	Attending meetings of leaders club.	D	M	Y	N
101K.	Attending meetings of a model airplane club	D	M	Y	N
102C.	Collecting scrap books.	D	M	Y	N
103H.	Taxidermy	D	M	Y	N
104X.	Skin diving	D	M	Y	N

ACTIVITIES REPORT (Continued)

105R.	Reading western stories	D	M	Y	N
106S.	Archery	D	M	Y	N
107M.	Taking music lessons.	D	M	Y	N
108K.	Working on school annual.	D	M	Y	N
109K.	Attending meetings of a camera club .	D	M	Y	N
110C.	Collecting colored pictures	D	M	Y	N
111H.	Mounting fish	D	M	Y	N
112X.	Operating a movie projector	D	M	Y	N
113R.	Reading airplane stories.	D	M	Y	N
114S.	Hunting	D	M	Y	N
115M.	Listening to records.	D	M	Y	N
116K.	Attending meetings of a forensics club	D	M	Y	N
117K.	Attending meetings of a Dairy Herd Improvement Association	D	M	Y	N
119H.	Banding birds	D	M	Y	N
120X.	Working on automobiles.	D	M	Y	N
121R.	Reading nonfiction.	D	M	Y	N
122S.	Fishing	D	M	Y	N
123M.	Listening to radio music.	D	M	Y	N
124K.	Attending meetings of the Junior Red Cross	D	M	Y	N

ACTIVITIES REPORT (Continued)

125K.	Attending meetings of a Horsemanship club	D	M	Y	N
126C.	Collecting cups	D	M	Y	N
127X.	Listening to radio or television humor programs.	D	M	Y	N
128R.	Reading classics	D	M	Y	N
129S.	Horseback riding	D	M	Y	N
130K.	Attending meetings of an Ushers club.	D	M	Y	N
131K.	Attending meetings of a Bird Club . .	D	M	Y	N
132C.	Attending meetings of natural objects. . . .	D	M	Y	N
133X.	Collecting to radio or television mysteries	D	M	Y	N
134R.	Reading war stories	D	M	Y	N
135S.	Roller skating.	D	M	Y	N
136K.	Attending meetings of a radio-television club.	D	M	Y	N
137K.	Attending meetings of a fly-casting club	D	M	Y	N
138C.	Collecting candles.	D	M	Y	N
139X.	Working puzzles	D	M	Y	N
140R.	Reading movie magazines	D	M	Y	N

ACTIVITIES REPORT (Continued)

141S.	Trapping	D	M	Y	N
142K.	Attending meetings of a Library Round Table	D	M	Y	N
143K.	Attending meetings of Civil Air Patrol	D	M	Y	N
144C.	Collecting plants	D	M	Y	N
145X.	Working crossword puzzles	D	M	Y	N
146R.	Reading historical novels	D	M	Y	N
147S.	Shooting	D	M	Y	N
148K.	Attending meetings of a dolphin club	D	M	Y	N
149K.	Serving as an officer of a club	D	M	Y	N
150C.	Collecting snapshots.	D	M	Y	N
151X.	Playing with young children	D	M	Y	N
152R.	Reading plays	D	M	Y	N
153S.	Skiing	D	M	Y	N
154K.	Attending meetings of a Boys' Hobby Club	D	M	Y	N
155C.	Collecting symbol pins	D	M	Y	N
156X.	Teaching children to dance	D	M	Y	N
157R.	Reading art books.	D	M	Y	N
158S.	Ice skating	D	M	Y	N
159K.	Attending meetings of a Nature Club	D	M	Y	N

ACTIVITIES REPORT (Continued)

160C.	Collecting pennants	D	M	Y	N
161X.	Traveling	D	M	Y	N
162R.	Reading books about mechanics	D	M	Y	N
163S.	Weight lifting	D	M	Y	N
165X.	Taking correspondence courses	D	M	Y	N
166S.	Playing on school sports team.	D	M	Y	N
167C.	Collecting paintings.	D	M	Y	N
168X.	Social dancing.	D	M	Y	N
169S.	Playing on out-of-school sports team	D	M	Y	N
170C.	Collecting china dolls.	D	M	Y	N
171X.	Tap dancing	D	M	Y	N
172C.	Collecting stuffed animals.	D	M	Y	N
173C.	Collecting butterflies	D	M	Y	N
174C.	Collecting snake skins.	D	M	Y	N
175X.	Ballet dancing	D	M	Y	N
176X.	Attending dance recitals.	D	M	Y	N
177X.	Daydreaming	D	M	Y	N
178X.	Watching television	D	M	Y	N
179X.	Going to movies	D	M	Y	N
180X.	Watching sports	D	M	Y	N
181X.	Going to parties	D	M	Y	N

ACTIVITIES REPORT (Continued)

182X.	Taking pictures	D	M	Y	N
183X.	Writing pen pals	D	M	Y	N
184X.	Just bumming around	D	M	Y	N
185X.	Going to church schools	D	M	Y	N
186X.	Watching musical entertainers	D	M	Y	N
187X.	Flying	D	M	Y	N
188X.	Going with a gang	D	M	Y	N
189X.	Gardening	D	M	Y	N
190X.	Just sitting around	D	M	Y	N
191X.	Hanging around the dairy bar	D	M	Y	N
192X.	Playing cards	D	M	Y	N
193X.	Meeting new people	D	M	Y	N
194X.	Playing chess	D	M	Y	N
195X.	Riding a bicycle	D	M	Y	N
196X.	Doodling	D	M	Y	N
197X.	Playing checkers	D	M	Y	N
198X.	Learning to drive a car.	D	M	Y	N
199X.	Working with machinery.	D	M	Y	N
200X.	Clerking in a store	D	M	Y	N

Other kinds of books or magazines you have read_____

ACTIVITIES REPORT (Continued)

Other sports you play_____

Other kinds of musical activities you engage in_____

Other clubs or other organizations you belong to_____

Other things you collect_____

Other things you do that are not given in any of the
lists above_____

Did you have a job during the past summer?_____

Part-time_____Full time_____How long?_____(Weeks)

Type of job_____Describe what you actually
did on the job_____

ACTIVITIES REPORT (Continued)

Did you like that kind of work? _____

Do you now have a part-time job? _____

How long have you had it? _____

Type of job _____ Describe what you actually
do on the job _____

Do you like the work? _____

If you don't have a job now, do you expect to get one? _____

What kind? _____

ANALYSIS OF REPORT ON ACTIVITIES

Name _____ Name of School _____
Location _____ Date _____ Birth Date _____
Grade _____ Interviewer _____

Those activities that appear, to the counselor, to occupy a significant portion of a counselor's time are reported below. Those marked with an asterisk(*) were checked as being very important by the student.

[illegible]

Supplementary Information and/or Counselor's Comments:

INDEX OF NAMES

- Abt, L. E., 318
 Adams, G., 62
 Allport, G. W., 227
 Anastasi, A., 47, 113, 126, 140, 149, 343, 351
 Anderson, G. L., 318, 321
 Anderson, G. V., 25
 Anderson, H. H., 318, 321
 Andrew, D. C., 62

 Barthol, R. P., 318
 Bayley, N., 142
 Belin, H., 248
 Bellak, L., 318
 Bennett, G. K., 25, 86, 100, 340, 341
 Benton, A. L., 293
 Berdie, R. F., 246, 278, 283
 Berg, I. A., 106
 Berkshire, J. R., 283
 Bittner, R. H., 113
 Bixler, R. H., 249, 269
 Bixler, V. H., 249, 269
 Bond, G. L., 114
 Bordin, E. S., 268, 269, 293
 Brewster, R. E., 347, 351
 Brookover, W. B., 244
 Brown, C. W., 113
 Buros, O. K., 36, 47, 149, 282, 314, 344, 351

 Caplow, T., 246, 280
 Carter, R. S., 237, 280
 Cattell, Jaques, 119
 Cattell, J. M., 119
 Cattell, P., 126
 Cattell, R. B., 289, 303
 Christenson, T. E., 41
 Coleman, J. C., 149
 Cook, W. W., 224
 Cottle, W. C., 107, 113
 Courtis, S. A., 224
 Cronbach, L. J., 25, 47, 63, 113, 149
 Crowder, N. A., 149
 Cruze, W. W., 63

 Cureton, E. E., 113, 149, 343

 Danielson, P. J., 6
 Darley, J. G., 25, 288, 290
 Davis, A., 342
 Dewey, J., 15
 Dolansky, M. P., 346
 Doppelt, J. E., 25, 149, 346
 Douglas, H. R., 237
 Drake, J., 113
 Dreger, R. M., 149
 Dressel, P. E., 269
 Duran, J. C., 90
 Durnall, E. J., Jr., 318
 Durrell, D. D., 133, 149
 Dvorak, B. J., 13, 327, 341

 Eells, K., 244, 280, 342
 Ellis, A., 318, 343, 351
 Englehart, M. D., 343, 346
 Eron, L. D., 318

 Faries, M., 268
 Flanagan, J. C., 64, 113, 224, 341
 Forbes, F., 107
 Frank, L. K., 288, 318
 Freeman, F. S., 47, 149, 343
 Fricke, B. G., 289
 Furst, E. J., 289

 Gardner, E. F., 224
 Garrett, H. E., 141, 149
 Gaylord, R. H., 113
 Gheselli, E. E., 113, 318
 Goodenough, F., 47
 Guilford, J. P., 128, 229, 232, 280, 318
 Gustad, J. W., 280

 Hagen, E., 150, 343, 352
 Hahn, M. E., 286, 287
 Hamlin, R. M., 319
 Havighurst, R. L., 247
 Heumann, R. A., 227, 243, 280, 289, 304, 308, 319

- Henmon, V. A. C., 340
 Herzberg, F. I., 88
 Hildreth, G. H., 47
 Hilkert, R. N., 96
 Hollingshead, A. B., 246, 280
 Hollingshead, B. S., 246, 280
 Hotelling, H., 128
 Hull, C. L., 232, 280, 325, 327
 Humphreys, J. A., 62
 Hunt, H. C., 251
- Jacobs, R., 287, 309
 Jenkins, J. G., 114
 Johnson, G. H., 298
 Johnson, R. H., 114
 Jones, A. J., 26, 63, 114, 347
 Jordan, A. M., 62
- Kelley, T. L., 128, 150, 341
 Kirk, B. A., 47, 268, 343
 Knezevich, S., 150
 Kornhauser, S. I., 293
 Kuder, G. F., 319, 351
- Laitin, Y. J., 47
 Learned, W. S., 237
 Lecky, P., 227
 Lennon, R. T., 114, 126, 134, 150, 346
 Leonard, W. N., 244, 280
 Lincoln, E. A., 290
 Lindquist, E. F., 42, 71, 79, 106, 113,
 114, 115, 150
 Lindzey, G., 319
 Loch, M. B., 247
 Longstaff, H. P., 293
 Lorge, I., 114
 Lyton, W. L., 309
- McCabe, G. E., 232, 280
 McCall, W. A., 224
 MacLean, M. S., 286, 287
 McNemar, Q., 245
 MacQuarrie, T. W., 83, 98
 Manuel, H. T., 114
 Mathewson, R. H., 26
 Matteson, R. W., 269
 Meehl, P. E., 113, 267, 280
 Merrill, M. A., 123
 Miller, L. M., 347
 Morse, W. C., 288
 Murphy, G., 227
 Murray, H. A., 321
- Nelson, M. J., 340
- Olson, N., 237
 Otis, A. S., 141, 150, 228
- Patterson, C. H., 114
 Patterson, D. S., 293
 Phearman, L. T., 246
 Pierce-Jones, J. A., 107
 Pollack, A. B., 114
 Prator, R., 251
- Rapaport, G. M., 106
 Raskin, E., 341
 Rath, L. E., 314
 Remmers, H. H., 114
 Ricks, J. H., Jr., 313
 Roebber, E. C., 297
 Roens, B. A., 4, 5, 9, 15, 26, 114, 238,
 252, 280, 307, 315, 334
 Rogers, C. R., 267, 282
 Ross, C. C., 36, 343, 351
 Ross, E. E., 119
 Rothney, J. W. M., 4, 5, 6, 9, 15, 26,
 99, 114, 150, 208, 224, 238, 248, 252,
 254, 268, 269, 280, 289, 304, 307,
 308, 315, 319, 334, 343, 345, 353
 Ruch, F. L., 25
 Rulon, P. J., 150, 224, 332, 351
- Sanderson, H. Z., 250
 Schenk, Q. L., 245, 280
 Schmidt, L. A., 150, 319, 343
 Seashore, H. G., 100, 313, 321, 329, 340
 Segal, D., 341
 Selby, P. O., 114
 Shaffer, L. T., 314
 Smith, E. R., 205, 224, 237, 280
 Sorokin, P. A., 331, 343, 351
 Spencer, L. M., 11
 Srole, L., 246
 Stanley, J. C., 36, 343, 351
 Steward, V., 293
 Stewart, N., 150
 Stordahl, K. E., 319
 Stoughton, R. W., 348
 Strang, R., 103
 Strong, E. K., 291, 319
 Stuit, D. B., 114
 Stunkel, E. R., 113
 Sullivan, E. T., 128
 Super, D. E., 17, 26, 47, 103, 114, 150,
 268, 290, 342, 343

- Terman, L. M., 123
Thorndike, R. L., 26, 67, 70, 75, 79, 96,
114, 150, 226, 280, 324, 343, 352
Thorpe, L. P., 63, 287
Thurstone, L. L., 128, 225, 228, 319
Thurstone, T. G., 150, 228
Tiedeman, D. V., 215
Tiegs, E. W., 127, 128, 129
Toops, H. A., 228
Torgerson T. L., 62
Traeger, C., 41
Travers R. M. W., 72, 115, 134, 150,
343, 352
Traxler, A. E., 47, 62, 78, 93, 96, 104,
105, 106, 115, 141, 150, 225, 288,
321
Trinkans, W. K., 309
Tyler, L. E., 268, 269, 286, 312
Tyler, R. W., 205, 224, 290
Warner, W. L., 246, 247
Warters, J., 26
Wesman, A. G., 100, 230, 234, 280,
324, 340
Whisler, L. D., 114
Whyte, W. H., Jr., 12, 26
Wilder, C. E., 113
Willey, J. M., 62
Windle, C. D., 319
Wingo, J. M., 288
Woellner, R. C., 268
Wolfe, D. L., 26, 246, 280

INDEX OF SUBJECTS

- Achievement, academic, 229, 234, 237-239, 244
 Activities reports, 353-370
 Alienation, coefficient of, 229
 American Council on Education Psychological Examination, 161, 228, 242, 263
 American Educational Research Association, 153, 342, 344, 345
 American Personnel and Guidance Association, 347, 348
 Ethical Practices Report, 347
 American Psychological Association, 40, 48, 342, 348
 Ethical Practices Report, 347
 American Textbook Publishers' Institute, 38
 Answer sheets, 105-107, 108, 109, 145-147
 Aptitude Tests for Occupations, 61

 Behavior description, 252-255
 Bennett Stenographic Aptitude Test, 46, 57
 Bennett Test of Mechanical Comprehension, 58, 84, 89, 116, 263
 Bernreuter Personality Inventory, 302
 Blyth Second-Year Algebra Test, 60
 Bureau of Publications, 39

 California Achievement Test, 242
 California Test Bureau, 38, 335
 California Test of Personality, 282, 302
 California Tests of Mental Maturity, 57, 107, 108, 127-130, 210, 261
 Cases, Art, 263-264
 Barbara, 263-266
 Bert, 258-260
 Bill, 264
 Bob, 260-262
 Dave, 262-263
 Ed, 255-258
 George, 271-273
 Jeff, 241-244
 Jerry, 236
 Mack, 263
 Mike, 273-279
 Raoul, 263
 Shelia, 239-241
 College Entrance Examination Board, 37
 Committee on Diagnostic Reading Tests, 38
 Computers, electronic, 328-329
 Confidence levels, 188
 Consistency, internal, 192-194
 of performance, 65
 trait, 227
 Coöperative Test Division, 38, 161, 343
 Counseling, objectives of, 4
 testing in, 14-16, 20-22, 336
 Counselors, certification of, 347-349
 questions asked by, 8
 role of, 5-6
 Cumulative records, 204

 Davis-Eells Games, 342
 Differential Aptitude Tests, 46, 68, 82, 84, 86, 87, 99, 100, 108, 112, 210, 230, 234, 242, 256, 259, 261, 263, 264, 277, 340, 341

 Educational Records Bureau, 37-38, 91, 338
 Educational Test Bureau, 38
 Educational Policies Commission, 324
 Educational Testing Service, 36-37, 107, 195
 Eight-Year Study, 212
 Employers, questions asked by, 8
 Expectancy tables, 216, 230-231

 False positives, 219
 Flanagan Aptitude Classification Tests, 37, 327-328, 341
 Follow-up studies, 214
 Forced-choice techniques, 284, 297

- General Aptitudes Test Battery, 40-42,
327, 341
- Grades, 237-239, 244
- Guidance workers, standards of, 347-350
- Harvard University Press, 39
- Henmon-Nelson Test of Mental Ability,
46, 107, 110, 245, 256, 259, 262,
277, 279, 340
- Heston Personal Adjustment Inventory,
302, 303
- Holzinger-Crowder Uni-Factor Tests, 57
- Houghton Mifflin Company, 39
- Information, sources of in counseling,
9-10
- Institutions for advanced training, ques-
tions asked by, 8
- Interest inventories, critical evaluation of,
282-301
norms of, 310-313
reliability of, 308-310
validity of, 290-301
- Interviews, 10, 282
- Intrapersonal factors and test scores,
248-252
- Iowa Algebra Aptitude Test, 58
- IQ, concepts of, 125-128
- K-scores, 215
- Kuder Preference Record, 56, 210, 282,
288, 293, 316, 318
- Kuhlman-Anderson Test, 239, 261
- Large-Thorndike Intelligence Tests, 57
- MacQuarrie Test of Mechanical Abili-
ties, 46, 84, 90, 98
- Manpower utilization, testing in, 16-20,
324-325
vs. human development concepts, 20-
21
- Measurement, actuarial concept of, 15
- Meier Art Test, 59
- Mental Measurement Yearbooks, 117,
148, 282, 314, 342, 344
- Minnesota Clerical Tests, 107
- Minnesota Multiphasic Personality In-
ventory, 210, 282, 301, 312
- Minnesota Paper Form Board, 84, 87,
107
- Motivation, and social class, 247-248
- effects of, on test performance, 248-
252
of pupils for testing, 99
- Multiple Aptitude Tests, 341
- Myers-Ruch High School Progress Test,
108
- National Council on Measurements Used
in Education, 93, 153, 342
- National Intelligence Tests, 107
- Norms, characteristics of, 83-85, 196-
202
in test interpretation, 81-82, 85-91,
135-138, 312-313, 338
local, 91-92
personality and interest inventory,
310-313
- Objectivity, 334-335
- Occupational Guide Series, 45
- Ohio State Psychological Test, 228, 239
- Otis Quick-Scoring Test of Mental Abil-
ity, 56, 98, 107, 126, 146, 228,
239, 242
- Parents, questions asked by, 7
- Percentile, in reporting to parents and
teachers, 81-83, 215
in test score interpretation, 208-209,
215-216
- Personality appraisal techniques, norms
of, 310-313
reliability of, 308-310
validity of, 301-303
- Pintner-Cunningham Test, 261
- Prediction, 50-52, 216-219, 228-236
- Primary Mental Abilities Tests, 46
- Probability, 232
- Profiles, as method of recording test re-
sults, 210-211
test interpretation from, 211, 311
- Progressive Achievement Tests, 46, 277
- Progressive Education Association, 252
- Projective methods, critical evaluation of,
303-308
reliability of, 308-310
validity of, 304-308
- Psychological Corporation, 39-40, 207,
344
- Test Service Bulletins of, 64, 208,
230, 233, 234, 346
- Public School Publishing Company, 38

- Reliability, 65-81, 187-192
 - and validity, 80-81
 - coefficients of, 67, 68, 72, 73, 110, 188, 308, 310
 - equivalent forms, 70-71
 - factors influencing, 71-75, 138-139
 - Kuder-Richardson reliability method, 70-71, 189, 190, 194
 - long-term, 81, 138, 141, 195
 - methods of determining, 66-70
 - of individual scores, 75-78
 - of speed tests, 67-69
 - split-half, 66-69, 73, 138, 308-309
 - test-retest, 69-70, 309-310
- Rogers Test of Personality Adjustment, 282
- Rorschach Inkblot Test, 282, 301
- School and College Ability Tests, 61, 74, 95, 107, 160-203, 341, 343
- School marks, 229, 234, 237-239, 244
- School personnel, questions asked by, 7-8
- Science Research Associates, 38
- Scores, scaled, 214
 - stability of, 78-80
 - true, 74-75
- Scoring machine vs. hand, 105-106
- Selection, and counseling, 235-237
 - testing for, 18-21
- Selective Service College Qualification Tests, 40
- Self-concept, 250-251, 289
- Sequential Tests of Educational Progress, 341, 343
- Short-cut methods, attempts to justify, 285-288
 - limitations of, 284-285
 - objections to, 288-290
- Social class status, 244-247
- SRA Clerical Test, 60
- SRA Mechanical Aptitudes Test, 84
- SRA Primary Mental Abilities Tests, 46, 56, 88, 108, 256, 259, 277
- Stability, coefficient of, 78-80, 309
- Standard error of estimate, 233
- Standard error of measurement, 76-78, 140, 232
- Stanford Achievement Tests, 210
- Stanford-Binet, 46, 123, 126, 320
- Stanford University Press, 39
- Statistical methods, 336-337
- Stenographic Aptitude Tests, 68
- Strong Vocational Interest Blank, 282, 286, 288, 291-292, 293, 311, 318
- Students, questions asked by, 7
- Studies, longitudinal, 227, 238, 336-337
- Subjectivity, 267, 334-335
- Survey of Mechanical Insight, 59
- T-scores, 214
- Teacher evaluations, 237-239, 240
- Technical Recommendations for Achievement Tests, 94, 342
- Technical Recommendations for Psychological Tests and Diagnostic Techniques, 48, 71, 93, 151-203
- Terman Group Test of Mental Ability, 124
- Terman-McNemar Test of Mental Ability, 123, 125
- Test administration, 92-104, 195-196
 - directions for, 97-98, 142-144
 - importance of, 92-95
 - motivation of students through, 97-100, 268
 - observations during, 103-104
 - physical conditions of, 96-97
- Test authors, 119-120
- Test coverage, 33, 131-132
- Test development, assumptions involved in, 123-131
 - basic problems in, 330-334
- Test evaluation, criteria for, 151-203
- Test improvement and consumer demand, 329-330, 343-347
- Test interpretation, implied validity in, 55-61, 131-135, 268-271
- Test materials, sources of, 36-42
- Test of Basic Skills in Arithmetic, 58
- Test publishers, 36-40, 120-122
 - commercial, 38-39
 - criteria for checking, 42-43
- Test results and contradictory evidence, written reports of, 255-266
- Test scores, 104-107, 144-147, 195-196
 - and other data, 265-271
 - cumulative recording of, 204-209, 213-216
 - etiology of, 116-147
 - factors influencing, 119, 226-228, 339
 - in counseling, 130, 133-135
 - prediction from, 216-219
 - profiles of, 210-211
 - recording and reporting of, 204-221

- Test selection, 27-36, 151-203
 - criteria of, 30-31
- Test titles vs. content, 33-36, 122-131, 295-296
- Testing, and free schools, 337-339.
 - differential, 340-342
 - future of, 339-350
 - in Europe, 18
 - in selection, 235-237
 - industrial, 18-19
 - scope of, 10-14
- Testing programs, mental ability tests
 - used in, 130-131
 - purposes of, 131-132
- Tests, apparatus, 28, 29
 - consumer interest in, 329-330
 - coöperative, 239, 241, 259, 263, 318
 - development of, 322-339
 - format of, 28-29
 - group, 28, 29
 - group vs. individual, 31
 - individual, 28
 - information provided by, 229
 - joint use of, 219-221
 - locally constructed, 42
 - mechanical considerations of, 107-109
 - miniature vs. trait, 28, 32
 - nonverbal, 28-29
 - power, 28-29
 - professional acceptance of, 62-65
 - recognition vs. demonstration, 33-34
 - reviews of, 344-345
 - single score vs. battery, 28, 31-32
 - speed, 28, 249, 335
 - standards for, 330-331
 - time-limit, 28, 29
 - trait, 28, 227
 - types of, 27-30
 - verbal, 28, 29
 - work-limit, 28, 29
- Thematic Apperception Test, 282, 288
- Thurstone Temperament Schedule, 302, 303
- Turse Clerical Aptitudes Test, 72
- United States Employment Service, 40, 327, 341
 - testing program, 40-42
- Validity, 48-64, 131-135, 174-187
 - coefficients of, 49-65
 - concurrent, 52-53, 180-186
 - construct, 53-54, 186-187
 - content, 50, 174-179
 - group vs. individual, 55
 - of personality and interest inventories, 301-308
 - predictive, 50-52, 57, 65, 179-180, 303, 331
- Van Wagenen Reading Readiness Test, 108
- Wechsler Adult Intelligence Scale, 242
- Wechsler Intelligence Scale for Children, 207, 208, 261, 263
- Wisconsin Counseling Study, 100
- World Book Company, 39, 346
- World War II, testing program in, 18
- Youth, problems of, 2-4
- Z-scores, 207